Transcriptome-wide association study reveals candidate causal genes for lung cancer

Yohan Bossé[1,2], Zhonglin Li[1], Jun Xia[3], Venkata Manem[1], Robert Carreras Torres[4], Aurélie Gabriel[4], Nathalie Gaudreault[1], Demetrius Albanes[5], Melinda C. Aldrich[6], Angeline Andrew[7], Susanne Arnold[8], Heike Bickeboller[9], Stig E. Bojesen[10], Paul Brennan[4], Hans Brunnstrom[11], Neil Caporaso[12], Chu Chen[13], David C. Christiani[14], John K. Field[15], Gary Goodman[16], Kjell Grankvist[17], Richard Houlston[18], Mattias Johansson[4], Mikael Johansson[19], Lambertus A. Kiemeney[20], Stephan Lam[21], Maria Teresa Landi[22], Philip Lazarus[23], Loic Le Marchand[24], Geoffrey Liu[25], Olle Melander[11], Gadi Rennert[26], Angela Risch[27], Susan M. Rosenberg[28,29], Matthew B. Schabath[30], Sanjay Shete[31], Zhuoyi Song[28,29], Victoria L. Stevens[32], Adonina Tardon[33], H-Erich Wichmann[34], Penella Woll[35], Shan Zienolddiny[36], Ma'en Obeidat[37], Wim Timens[38], Rayjean J. Hung[39], Philippe Joubert[1], Christopher I. Amos[3], James D. McKay[4]

1) Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Quebec City, Canada
2) Department of Molecular Medicine, Laval University, Quebec City, Canada
3) The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA
4) International Agency for Research on Cancer, World Health Organization, Lyon, France
5) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA
6) Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, USA
7) Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, USA
8) Markey Cancer Center, University of Kentucky, Lexington, USA
9) Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, Germany
10) Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark
11) Clinical Sciences, Lund University, Lund, Sweden
12) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, USA
13) Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
14) Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA
15) Molecular and Clinical Cancer Medicine, Roy Castle Lung Cancer Research Programme, The University of Liverpool Institute of Translational Medicine, Liverpool, UK
16) Public Health Sciences Division, Swedish Cancer Institute, Seattle, USA
17) Department of Medical Biosciences, Umeå University, Umea, Sweden
18) German Research Center for Environmental Health, Institute for Cancer Research, London, United Kingdom
19) Department of Radiation Sciences, Umeå University, Umea, Sweden
20) Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, Netherlands
21) Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, Canada
22) National Cancer Institute, Bethesda, USA
23) College of Pharmacy, Washington State University, Spokane, USA
24) Epidemiology, University of Hawaii Cancer Center, Honolulu, USA
25) Epidemiology Division, Princess Margaret Cancer Center, Toronto, Canada

26) Technion Faculty of Medicine, Carmel Medical Center, Haifa, Israel
27) Cancer Center Cluster Salzburg at PLUS, Department of Molecular Biology, University of Salzburg, Heidelberg, Austria
28) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
29) Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA
30) Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, USA
31) Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, USA
32) Epidemiology Research Program, American Cancer Society, Atlanta, USA
33) Faculty of Medicine, University of Oviedo and CIBERESP, Oviedo, Spain
34) Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig Maximilians University, Munich, Germany
35) Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, UK
36) National Institute of Occupational Health (STAMI), Oslo, Norway
37) The University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada
38) University of Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, GRIAC Research Institute, Groningen, the Netherlands
39) Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

Running title
TWAS on lung cancer

Keywords
Lung cancer, transcriptome-wide association study, GWAS, lung eQTL

Corresponding author
Yohan Bossé, Ph.D.
Professor, Laval University
Department of Molecular Medicine
Canada Research Chair in Genomics of Heart and Lung Diseases
Institut universitaire de cardiologie et de pneumologie de Québec
Pavillon Marguerite-d'Youville, Y2106
2725, chemin Ste-Foy,
Québec (Québec),
Canada, G1V 4G5
Tel: 418-656-8711 ext. 3725
Fax: 418-656-4940
email: yohan.bosse@criucpq.ulaval.ca

Conflicts of interest
The authors disclose no potential conflicts of interest.

Word count: 5273
Total Number of Figures and Tables: 2 tables and 5 figures

Novelty & Impact Statements

Genome-wide integration of GWAS and lung eQTL reveals candidate target genes for lung cancer. These include a new susceptibility locus for lung adenocarcinoma on chromosome 9p13.3 with *AQP3* (aquaporin 3) as the most likely target gene and *IREB2* (iron responsive element binding protein 2) on chromosome 15q25 for all histological subtypes. Putative causal genes acting through expression in lung provide insights about disease etiology and refine the biological interpretation of previous GWAS.

Abbreviations

COPD, chronic obstructive pulmonary disease
dbGaP, database of Genotypes and Phenotypes
enet, elastic net regression
eQTL, expression quantitative trait loci
FDR, false-discovery rate
GTEx, Genotype-Tissue Expression
GWAS, genome-wide association study
LASSO, least absolute shrinkage and selection operator
lncRNA, long non-coding RNA
MHC, major histocompatibility complex
SCLC, small cell lung cancer
TRICL-ILCCO, Transdisciplinary Research in Cancer of Lung team of the International Lung Cancer Consortium
TWAS, transcriptome-wide association study

Abstract

We have recently completed the largest GWAS on lung cancer including 29,266 cases and 56,450 controls of European descent. The goal of this study has been to integrate the complete GWAS results with a large-scale expression quantitative trait loci (eQTL) mapping study in human lung tissues (n=1,038) to identify candidate causal genes for lung cancer. We performed transcriptome-wide association study (TWAS) for lung cancer overall, by histology (adenocarcinoma, squamous cell carcinoma, small cell lung cancer) and smoking subgroups (never- and ever-smokers). We performed replication analysis using lung data from the Genotype-Tissue Expression (GTEx) project. DNA damage assays were performed in human lung fibroblasts for selected TWAS genes. As expected, the main TWAS signal for all histological subtypes and ever-smokers was on chromosome 15q25. The gene most strongly associated with lung cancer at this locus using the TWAS approach was *IREB2* (P_{TWAS}=1.09E-99), where lower predicted expression increased lung cancer risk. A new lung adenocarcinoma susceptibility locus was revealed on 9p13.3 and associated with higher predicted expression of *AQP3* (P_{TWAS}=3.72E-6). Among the 45 previously described lung cancer GWAS loci, we mapped candidate target gene for 17 of them. The association *AQP3*-adenocarcinoma on 9p13.3 was replicated using GTEx (P_{TWAS}=6.55E-5). Consistent with the effect of risk alleles on gene expression levels, *IREB2* knockdown and *AQP3* overproduction promote endogenous DNA damage. These findings indicate genes whose expression in lung tissue directly influence lung cancer risk.

Introduction

Genome-wide association studies (GWAS) to date have reported 45 lung cancer susceptibility loci in European and Asian populations[1]. Identifying the causal genes underpinning these loci remains a major challenge. Expression quantitative trait loci (eQTL) in disease relevant tissues are known to complement GWAS results by providing the specific genes whose expression levels are associated with disease-associated SNPs[2]. This strategy has been applied in lung cancer by directly testing disease-associated SNPs for association with expression levels of nearby genes in lung tissues[3].

Recent development in bioinformatics now allows transcriptome-wide association study (TWAS), which is a more advanced approach to integrate GWAS and eQTL results and identify candidate causal genes underlying diseases[4, 5]. TWAS requires a set of individuals for whom both gene expression and genetic variants have been measured, i.e. an eQTL dataset. The part of gene expression that can be explained by *cis*-acting SNPs can then be modeled in the eQTL dataset and used to impute the genetic component of expression in a second (usually larger) set of individuals with only SNP GWAS data. The approach can be conceptualized as having imputed expression data for all cases and controls used in a GWAS without directly measuring expression levels in these samples. The association between imputed gene expression and the disease (or biological trait) of interest is then evaluated by performing a TWAS.

In this study, we combined the largest GWAS on lung cancer[6] and the largest lung eQTL study[7] to perform a TWAS on lung cancer, histological subtypes and smoking subgroups. The objective is to identify candidate target genes for lung cancer residing within and outside GWAS-nominated loci.

Materials and Methods

Lung eQTL dataset

The lung eQTL dataset consists of whole-genome genotyping (Illumina Human1M-Duo BeadChip) and gene expression (Affymetrix) in non-tumor lung tissues from patients who underwent lung surgery at three academic sites, Laval University, University of British Columbia, and University of Groningen, henceforth referred to as Laval, UBC, and Groningen, respectively. All lung specimens from Laval were obtained from patients undergoing lung cancer surgery and were harvested from a site distant from the tumor. At UBC, the majority of samples were from patients undergoing resection of small peripheral lung lesions. Additional samples were from autopsy and at the time of lung transplantation. At Groningen, the lung specimens were obtained at surgery from patients with various lung diseases, including patients undergoing therapeutic resection for lung tumors, harvested from a site distant from the tumor, and lung transplantation. Lung tissue processing and storage, DNA and RNA extraction, genotyping, microarray-based gene expression and lung *cis*-eQTL analyses have been described previously[7, 8]. Following standard microarray and genotyping quality controls, data on 1,038 patients were available. At Laval and UBC, written informed consent was obtained from all subjects and the study was approved by their respective ethics committee. At Groningen, lung specimens were provided by the local tissue bank of the Department of Pathology and the study protocol was consistent with the Research Code of the University Medical Center Groningen and Dutch national ethical and professional guidelines ("Code of conduct; Dutch federation of biomedical scientific societies"; http://www.federa.org).

GWAS dataset

The GWAS data were derived from the Transdisciplinary Research in Cancer of Lung team of the International Lung Cancer Consortium (TRICL-ILCCO) OncoArray project comprising 29,266 lung cancer cases and 56,450 controls of European ancestry based on OncoArray and other Illumina genome-wide arrays[6]. The GWAS was performed using logistic regression to evaluate the association of genetic variants with overall lung cancer and the predominant histological subtypes including adenocarcinoma (n=11,273), squamous cell carcinoma (n=7,426), and small cell lung cancer (SCLC) (n=2,664). Genetic variants were also tested for association in never- (n=2,355) and ever-smokers (n=23,223). For this study, summary statistics were available for more than 10 million genotyped and imputed SNPs for overall lung cancer, histological subtypes and smoking subgroups (range 10,333,102 to 11,268,805). All participating studies in the TRICL-ILCCO OncoArray project were approved by their local ethics committee and all participants signed an informed consent.

Transcriptome-wide association study

The TWAS was performed for lung cancer overall, histological subtypes, and smoking subgroups using two approaches, i.e. S-PrediXcan[5] and FUSION[4]. The lung eQTL dataset was used as the training set to derive the expression weights. Gene expression normalized for age, sex and smoking status from Laval, UBC, and Groningen were combined with ComBat[9].

For analysis with S-PrediXcan, gene expression traits were first trained with elastic net linear models (alpha=0.5, n_k_folds=10, window=500 Kb) using the lung eQTL set. Models with false-discovery rate (FDR)<0.05 as implemented in S-PrediXcan were obtained for 19,889 probe sets. Predicted expression levels from the lung in the TRICL-ILCCO OncoArray project were then tested for association with lung cancer[5].

For analysis with FUSION, expression prediction models for each gene were evaluated in *cis*, using markers within 500 Kb on both sides of the expression probe sets. Probe sets that passed QC in the lung eQTL dataset (n=41,738) were evaluated and significant *cis*-heritability (P<0.01) were observed for 12,587 annotated probe sets. The best performing prediction models implemented in FUSION were LASSO regression and elastic net regression (enet) for 8,254 and 4,333 probe sets, respectively. Once the expression weights were obtained, expression imputation was performed using the summary statistics from the TRICL-ILCCO OncoArray project.
For both approaches, genome-wide significant TWAS were considered at $P_{TWAS}<0.05$ based on Bonferroni correction (S-PrediXcan $P_{TWAS}=0.05/19,889=2.51E-6$; FUSION $P_{TWAS}=0.05/12,587=3.97E-6$). A more liberal significant threshold was also used ($P_{TWAS}<0.0001$) to explore the top TWAS signals not reaching genome-wide significance for some histological or smoking subgroups. Finally, the top TWAS genes in previously established lung cancer risk loci that showed some evidence of association ($P_{TWAS}<0.05$) with both S-PrediXcan and FUSION were considered. For both TWAS approaches, we reported well-annotated probe sets. LocusCompare[10] was used to visualize GWAS and eQTL colocalization events.

Published GWAS loci for lung cancer

Lung cancer GWAS loci were derived from our recent review[1]. The boundaries of each locus were defined by adding 1 Mb downstream and upstream of lung cancer-associated SNPs derived from published GWAS on lung cancer. The genomic locations of TWAS genes were then overlapped with these lung cancer loci to delineate those residing within or outside GWAS loci.

TWAS replication

Lung eQTL data from 383 individuals available in the Genotype-Tissue Expression (GTEx) project (GTEx, version 7)[11] were used for TWAS replication. The TWAS was performed using S-PrediXcan and FUSION as described above.

In vitro assays

Cell line, plasmids and reagents. MRC-5V2 (male, SV40-immortalized human lung fibroblasts, Research Resource Identifier (RRID): CVCL_2627, source: Stephen P. Jackson Lab) cell line was maintained in Dulbecco's modified Eagle's medium (DMEM) (Gibco, Catalog #: 41965) supplemented with 10% fetal bovine serum (Gibco, Catalog #: 10438034), 2 mM L-glutamine, 100µg/mL penicillin, 100 µg/mL and streptomycin as previously described[12]. The human cell line has been authenticated using STR profiling within the last three years and all experiments were performed with mycoplasma-free cells. Gateway compatible *AQP3* entry clone was obtained from ccsbBroad gene libraries (ccsbBroadEn_00089). We subcloned *AQP3* into a mammalian expression vector containing a GFP epitope tag (pcDNA6.2/N-EmGFP-DEST, Invitrogen), which allows us to separate the transfected and non-transfected cell populations. Overproduction plasmids transfections were performed using GenJet (SignaGen, Catalog #: SL100488). SMARTpool *IREB2* and *NEXN* siRNAs as well as non-targeting (NT) siRNA were purchased from Dharmacon. siRNA transfections were carried out with lipofectamine RNAiMax (Invitrogen #13778150) following the manufacturer's instructions. Knockdown efficiency was evaluated by real-time quantitative reverse transcription PCR (qRT-PCR). RNeasy mini kit (Qiagen #74106) was used to extract from MRC-5V2 cells that were transfected with siRNA for 72 hours. 300 ng of total RNA from each sample was used to synthesize cDNA by the Superscript III first strand synthesis system (Invitrogen, #18080051). qPCR reactions were

performed using iTaq Universal SYBR Green Supermix (BioRad #172-5121). qPCR experiments were performed on the QuantStudio 3 Real-Time PCR System (Applied Biosystems). For each gene, three replicates were analyzed and the average threshold cycle (Ct) was calculated. The relative expression levels were calculated with the $2^{-\Delta\Delta Ct}$ method[13]. Primers used included *IREB2* forward: TCTTGGTATTACAAAGCACCTCAG; *IREB2* reverse: TCACATTGTCAACAGGGAAAAAG; *GADPH* forward: CAATGACCCCTTCATTGACC; *GADPH* reverse: GATCTCGCTCCTGGAAGATG; *NEXN* forward: ACTGTGAAGGGTAGATTTGCTG; *NEXN* reverse: TTCTGCGTTTTCGTTCCTCCT. Knockdown efficiency was 88% for *IREB2* and 95% for *NEXN*.

DNA damage assays by flow cytometry. Flow-cytometric DNA damage assays and quantification signals were performed as previously described[12]. Briefly, cells were fixed, permeabilized and stained with γH2AX antibody (Sigma, Catalog #05-636), then samples were measured by a BD LSRFortessa flow cytometer and analyzed using FlowJo software. For overproduction experiments, cells with mock transfection were used to set the threshold gating to determine the percentage of GFP^- and $\gamma H2AX^-$ cells, with 0.5% of control cells gated as the damage threshold as previously validated. The DNA-damage ratio caused by protein overproduction is defined by (Q2/Q3)/(Q1/Q4), where Q2 is the number of transfected damage-positive cells; Q3 is the number of transfected damage-negative cells; Q1 is the number of untransfected damage positive cells, and Q4 is the number of untransfected damage-negative cells.

Data availability

The GWAS data set used for the current study is available at the database of Genotypes and Phenotypes (dbGaP) under accession phs001273.v1.p1. The human lung tissue eQTL study is available in dbGaP under accession phs001745.v1.p1.

Results

Genes with *cis*-genetic component of expression in the lung

A total of 1,038 individuals for whom both gene expression and genetic variants were measured (i.e. the lung eQTL dataset) were used to impute the *cis* genetic component of expression into the larger set of 29,266 cases and 56,450 controls from the TRICL-ILCCO OncoArray project using their SNP genotype data (i.e. GWAS data). Integration of the lung eQTL and lung cancer GWAS was performed by two TWAS approaches, namely S-PrediXcan and FUSION. To be assessed by TWAS, a significant portion of gene expression had to be explained by SNPs. For S-PrediXcan, expression prediction models were obtained for 19,889 probe sets. On average, SNPs explained 4.95% of the probe sets expression variance, including 62.2% of probe sets that showed a prediction performance (R^2) of at least 0.01 (**Supplementary Figure 1A**). For FUSION, significant *cis*-heritability was observed for 12,587 annotated probe sets. On average, SNPs explained 7.39% of the probe sets expression variance, including 80.4% of probe sets for which their expression variance is explained by more than 1% (**Supplementary Figure 1B**). Significant *cis*-heritability was observed for 12,099 probe sets in both S-PrediXcan and FUSION (**Supplementary Figure 1C**) and the expression variance explained by SNPs for these probe sets was tightly correlated between the two methods (**Supplementary Figure 1D**).

Overall lung cancer

The TWAS results for overall lung cancer are illustrated in **Figure 1A**. TWAS genes that are statistically significant after Bonferroni correction are indicated in **Table 1**. The top TWAS signal is on chromosome 15q25, which is well-established as the strongest lung cancer susceptibility locus derived from GWAS. Interestingly, *IREB2* is the lead TWAS target gene on 15q25 by S-PrediXcan (P_{TWAS}=1.09E-99). Other statistically significant TWAS genes include

CHRNA3 (P_{TWAS}=4.66E-65), *CHRNA5* (P_{TWAS}=6.01E-22), *HYKK* (P_{TWAS}=6.57E-17), and *PSMA4* (P_{TWAS}=1.42E-13). In FUSION, *IREB2* also has a level of significance stronger (P_{TWAS}=4.97E-104) than other significant TWAS genes at this locus including *CHRNA5* (P_{TWAS}=5.26E-20), *HYKK* (P_{TWAS}=2.04E-17), and *PSMA4* (P_{TWAS}=4.15E-13). Lower predicted expression of *IREB2* is associated with increased lung cancer risk. **Figure 2** shows the colocalization of the GWAS and lung eQTL signals on 15q25 as well as the effect of the top GWAS SNP on the expression of *IREB2*. LocusCompare plots show the colocalization events for *IREB2* as well as other significant TWAS genes on 15q25 (**Supplementary Figure 2**). The lung cancer risk allele is associated with lowered expression of *IREB2* in lung tissues.

Significant TWAS genes are also identified at two loci on chromosome 6. The most significant, and containing the largest number of TWAS genes, is the MHC locus, including 23 significant genes (**Table 1**). The top TWAS gene is *APOM* in both S-PrediXcan and FUSION. In the MHC locus, lower predicted expression of 16 genes and higher predicted expression of 7 genes are associated with increased lung cancer risk. The direction of effect is consistent for the six genes in common between S-PrediXcan and FUSION (**Supplementary Figure 3**). The second locus on chromosome 6 (6q27) identifies *RNASET2* and *FGFR1OP* as the TWAS gene in S-PrediXcan (P_{TWAS}=2.33E-8) and FUSION (P_{TWAS}=7.68E-8), respectively.

Significant genes are observed at three additional loci. First, *RAD52* on 12p13.33 (P_{TWAS}=6.58E-10) by S-PrediXcan with higher predicted expression associated with higher lung cancer risk. Second, *SECISBP2L* on 15q21.1 by S-PrediXcan (P_{TWAS}=5.44E-9) and FUSION (P_{TWAS}=8.01E-10), which we have recently identified as the candidate target gene[6]. Third, *JAML*

on 11q23.3 by S-PrediXcan (P_{TWAS}=2.64E-7) and FUSION (P_{TWAS}=1.39E-6) with lower predicted expression associated with higher lung cancer risk.

Overall, TWAS genes are identified in six lung cancer susceptibility loci previously established by GWAS (**Table 1**). A potentially novel susceptibility gene is identified for 6q27-*FGFR1OP*. For the other five loci, the TWAS results refined putative causal genes suspected by GWAS and demonstrated their direction of effects with lung cancer risk. LocusCompare plots for these TWAS hits are provided in **Supplementary Figure 4**.

Histological subtypes

TWAS results by histological subtypes are shown in **Figures 1B-D** and **Table 1**. *IREB2* is the top TWAS gene for the three predominant subtypes, namely adenocarcinoma, squamous cell carcinoma and SCLC. Consistent with overall lung cancer, lower predicted expression is associated with increased risk of all histological subtypes.

For **adenocarcinoma**, consistent results between S-PrediXcan and FUSION are observed for *NRG1* on 8p12 (S-PrediXcan P_{TWAS}=3.29E-8, FUSION P_{TWAS}=1.21E-7) and *AQP3* on 9p13.3 (S-PrediXcan P_{TWAS}=3.72E-6, FUSION P_{TWAS}=3.49E-6). The latter is a new lung cancer susceptibility locus. **Figure 3** and **Supplementary Figure 5** show the colocalization of the GWAS and lung eQTL signals on 9p13.3 as well as the effect of the top GWAS SNP on the expression of *AQP3*. The lung cancer risk allele is associated with higher expression of *AQP3* in lung tissues. Additional TWAS genes for adenocarcinoma identified by S-PrediXcan and FUSION include *SECISBP2L* on 15q21.1 (S-PrediXcan P_{TWAS}=1.92E-16 and FUSION P_{TWAS}=2.50E-17), *TP63* on 3q28 (S-PrediXcan P_{TWAS}=2.50E-11 and FUSION P_{TWAS}=3.35E-12),

and *JAML* on 11q23.3 (S-PrediXcan P_{TWAS}=1.21E-8 and FUSION P_{TWAS}=2.09E-8). S-PrediXcan identifies *DCBLD1* on 6q22.1 (P_{TWAS}=3.59E-7). Lower predicted expression of all these genes (*DCBLD1*, *TP63*, *SECISBP2L*, *JAML*) is associated with increased risk of adenocarcinoma. All these loci were associated with lung cancer before. Interestingly, no significant TWAS gene in the MHC region was observed for adenocarcinoma.

For **squamous cell carcinoma**, the MHC region includes many TWAS genes (**Figure 1C** and **Table 1**). Similar to results observed for overall lung cancer, the top TWAS gene using S-PrediXcan and FUSION is *APOM*. There is one additional TWAS gene for squamous cell carcinoma by S-PrediXcan on 12p13.33. The target gene is *RAD52* (P_{TWAS}=1.24E-10) and the direction of effect indicates that higher expression is associated with an increased risk of squamous cell carcinoma. In FUSION, one more TWAS gene is identified for squamous cell carcinoma, namely *BLOC1S2* (P_{TWAS}=2.16E-6) on 10q24.31 with lower predicted expression associated with squamous cell carcinoma. *BLOC1S2* is a new candidate causal gene for squamous cell carcinoma.

For **SCLC**, the only significant TWAS gene other than *IREB2* and *CHRNA3* on 15q25 was *HIST1H2BD* on 6p22.2 (MHC locus) by FUSION (P_{TWAS}=1.54E-6) with predicted expression positively associated with SCLC. A second TWAS gene that just missed genome-wide significance is *TMA16* (P_{TWAS}=4.2E-6) on 4q32.2, which is a locus not yet reported for lung cancer. Higher predicted expression of *TMA16* is associated with higher risk of SCLC. S-PrediXcan did not provide a significant gene expression model for *TMA16*.

Smoking subgroups

TWAS results for ever- and never-smokers are in **Figures 1E-F** and **Table 1**. The TWAS in ever-smokers parallel results observed for overall lung cancer, albeit at lower significance levels. This includes *IREB2*, *CHRNA3*, and *CHRNA5* on 15q25, *SECISBP2L* on 15q21.1, *RAD52* on 12p13.33 by S-PrediXcan, and many genes in the MHC locus. The direction of effects is also consistent with overall lung cancer. For never-smokers, no TWAS gene reach genome-wide significance. One gene is identified using a more liberal significant threshold ($P_{TWAS}<0.0001$) using both TWAS approaches, namely *NEXN* on 1p31.1 with predict expression negatively associated with lung cancer in never-smokers. **Figure 4** shows the GWAS results for never-smokers on 1p31.1 and lung eQTL signals for *NEXN*. Colocalization events can further be visualized in **Supplementary Figure 6**. The lung cancer risk allele is associated with lower expression of *NEXN* in lung tissues. *NEXN* has never been reported as a lung cancer susceptibility gene.

Lung cancer risk loci from GWAS

We also explored the top TWAS genes in known lung cancer risk loci derived from previous GWASs. The boundaries of each locus were defined (see methods) and the top TWAS genes by S-PrediXcan and FUSION for overall lung cancer are indicated in **Table 2**. The top TWAS gene ($P_{TWAS}<0.05$) is consistent for both S-PrediXcan and FUSION at 6 additional loci (not in **Table 1**): *ORMDL1* on 2q32.2, *SLC22A5* on 5q31, *TRIM38* on 6p22.2, *MTAP* on 9p21.3, *N4BP2L2* on 13q13.1, and *MTMR3* on 22q12.2. Colocalization of GWAS and lung eQTL signals supports *ORMDL1*, *SLC22A5*, and *TRIM38* as candidate causal genes at these loci (**Supplementary Figure 7**). In contrast, the strongest lung eQTL variants for *MTAP*, *N4BP2L2*, and *MTMR3* have weak GWAS p values (**Supplementary Figure 7**), suggesting the possibility of false-positive

TWAS genes and the need to use alternative approaches to find the causal genes at these loci. Overall, we map candidate causal genes for 17 out of the 45 known lung cancer GWAS loci. **Supplementary Figure 8** summarizes candidate target genes for lung cancer identified in this study residing within and outside GWAS-nominated loci.

Replication in GTEx

The lung eQTL data from 383 individuals available in GTEx was used to validate the results. We first evaluated the new adenocarcinoma locus on 9p13.3-*AQP3*. The association *AQP3*-adenocarcinoma is strongly validated in GTEx (S-PrediXcan P_{TWAS}=6.55E-5 and FUSION P_{TWAS}=1.72E-5) with a consistent direction of effect, i.e. the risk allele increases the expression levels of *AQP3* in lung tissues. Second, we assessed *NEXN* as the new target gene underlying the 1p31.1 locus in never-smokers. The association and direction of effect were replicated (S-PrediXcan P_{TWAS}=0.006 and FUSION P_{TWAS}=0.003) with predicted expression negatively associated with lung cancer in never-smokers.

We also compared candidate target genes identified in GWAS-nominated loci. Note that replication of S-PrediXcan and FUSION results in GTEx lung data is only feasible for genes with significant prediction models. The sample size available for building lung models in GTEx is smaller (n=383) compared to our lung eQTL dataset (n=1,038). Therefore, replication is not feasible for a fraction of genes in GTEx lung, i.e. some genes will have no significant prediction model. This is the case for *IREB2* on 15q25 that did not yield a prediction model in GTEx lung. The top TWAS gene on 15q25 for overall lung cancer in GTEx lung is *CHRNA5* (P_{TWAS}=1.70E-14). Replication of all Bonferroni-corrected TWAS genes by histology and smoking subgroups is indicated in **Table 1**. Excluding the 15q25 and 6p-MHC loci, replication of TWAS genes was

observed for 3 out of 4 for overall lung cancer, 6 out of 6 for adenocarcinoma, 2 out of 2 for squamous cell carcinoma, 1 out of 1 for SCLC, 2 out of 2 for ever-smokers, and 1 out of 1 for never-smokers. Among the 6 additional loci showing the same top TWAS gene for both S-PrediXcan and FUSION, 4 could be evaluated in GTEx and 3 were replicated: 5q31-*SLC22A5*, 9p21.3-*MTAP* and 22q12.2-*MTMR3* (**Table 2**). Overall, for the 17 TWAS genes located in the 45 GWAS-nominated loci, 14 could be evaluated in GTEx and 12 were replicated.

Endogenous DNA damage assays

We hypothesized that some of the TWAS-nominated genes might promote cancer by increasing endogenous DNA damage, and subsequently lead to genome instability. Three TWAS genes were selected for *in vitro* assays: *IREB2* on 15q25, *AQP3* on 9p13.3, and *NEXN* on 1p31.1. The choice between knockdown and overproduction assays was guided by the direction of effect observed in the TWAS. For *IREB2* and *NEXN*, knockdown assays were performed to corroborate lower predicted expression associate with increased lung cancer risk, whereas overproduction assays were performed for AQP3 to mimic higher predicted expression associate with increased risk of lung cancer. We discovered that knockdown of *IREB2* increased endogenous DNA damage in human lung fibroblasts (**Figure 5**). In contrast, knockdown of *NEXN* had no effect on DNA damage. For AQP3, overproduction promotes endogenous DNA damage in lung fibroblasts (**Figure 5**).

Discussion

This study is the largest lung tissue based TWAS on lung cancer; gene expression prediction models built with a lung eQTL dataset of 1,038 individuals and association analyses of predicted gene expression with lung cancer risk using summary statistics derived from a GWAS on 29,266 cases and 56,450 controls. We revealed a new lung adenocarcinoma locus on 9p13.3 associated with the expression levels of *AQP3* in lung tissues. We also identified candidate causal genes at GWAS-nominated lung cancer loci including *IREB2* on 15q25 for all histological subtypes. Cellular DNA damage assays further supported the potential causality of lower predicted expression of *IREB2* and higher predicted expression of *AQP3* in increasing the risk of lung cancer. Overall, we map putative causal genes for 17 out of the 45 known lung cancer risk loci derived from GWAS.

During the last 10 years, GWAS have identified 45 susceptibility loci for lung cancer[1]. The genes underlying these genetic associations are largely unknown. As with other complex diseases, the GWAS risk variants for lung cancer are mostly located in non-coding regions and are thus believed to mediate their effects by influencing gene expression of nearby genes. In this study, we used a TWAS approach that captures the aggregate effects of multiple SNPs on gene expression and then tested the association of genetically predicted gene expression and disease risk. As a gene-based strategy, TWAS has the ability to identify the most likely target genes residing within GWAS-nominated loci, and also to reveal novel risk loci by the resulting power of combining GWAS and eQTL results. In this study, TWAS was performed using two competing approaches, i.e. S-PrediXcan and FUSION. Both belong to the same family of methods to discover gene-trait associations using models trained in eQTL datasets and summary-level GWAS data. The difference lies in the prediction models, i.e. S-PrediXcan uses elastic net

(enet), while FUSION evaluates different prediction schemes (herein: enet, LASSO, top1) and selects the best performing model. Using default parameters, we obtained more expression prediction models in S-PrediXcan compared to FUSION (19,889 vs. 12,587 probe sets with significant cis-heritability). However, the prediction performance of the 12,099 probe sets in common between S-PrediXcan and FUSION were tightly correlated, even when different prediction models (enet vs LASSO) were used (**Supplementary Figure 1D**).

The majority of TWAS genes identified in this study lie around known GWAS loci. The only SNP-level sub-genome-wide significant locus that yield genome-wide significant results by TWAS is 9p13.3-*AQP3* for adenocarcinoma. This novel susceptibility locus for adenocarcinoma (9p13.3-*AQP3*) was observed in S-PrediXcan and FUSION, and was also replicated in GTEx lung (**Table 1**). The direction of effect indicates that higher *AQP3* expression is associated with an increased risk of lung adenocarcinoma. *AQP3* (aquaporin 3) encodes a water channel protein that is expressed in the normal respiratory track and up-regulated in NSCLC, especially adenocarcinoma[14, 15]. Knockdown of *AQP3* has been shown to suppress proliferation and invasion of lung cancer cells[16, 17] as well as to inhibit tumor growth in human NSCLC xenografts[18]. The direction of effect observed in our study is thus concordant with these functional studies. In the current study, we further demonstrated that AQP3 overproduction promotes endogenous DNA damage in human lung fibroblasts. All together these observations support *AQP3* as the causal gene for lung adenocarcinoma on 9p13.3. The genetic association between *AQP3* and lung adenocarcinoma will require further validation.

Novel susceptibility genes were identified in previously established GWAS loci. In never-smokers, we have identified *NEXN* (nexilin F-actin binding protein) as the putative causal gene

on 1p31.1. Nexilin is an actin-binding protein known to play a role in cell adhesion and migration. Mutations in this gene have been associated with cardiomyopathy[19, 20]. The 1p31.1 locus was first demonstrated to be associated with lung cancer as part of a genome-wide investigation of SNPs within all long non-coding RNA (lncRNA) genes[21]. SNP rs114020893 located in lncRNA *NEXN-AS1* was associated with lung cancer and with similar association between adenocarcinoma and squamous cell carcinoma subgroups. *In silico* analysis then predicted that rs114020893 could change the folding structure of *NEXN-AS1*. However, it was unclear if the lung cancer-associated SNP was acting through *NEXN-AS1* or by regulating the expression of its corresponding gene, *NEXN*. Our current study supports the later. In this lncRNA study[21], analyses by smoking subgroups were not performed. In McKay et al.[6], the 1p31.1 locus was GWAS significant for overall lung cancer as well as for adenocarcinoma and ever-smoker subgroups, but did not reach significance in never-smokers. By using a TWAS approach, we demonstrated that this locus might also be relevant for the development of lung cancer in never-smokers and, at least in this subgroup, the susceptibility locus may mediate its effect by down-regulating the expression of *NEXN* in lung tissues. Interestingly, a recent study demonstrated that the expression levels of *NEXN-AS1* and *NEXN* are decreased in human atherosclerotic plaques and *NEXN* deficiency promotes atherosclerosis in an experimental mouse model[22]. *NEXN* seems to confer protection against atherosclerosis by suppressing inflammatory cytokines (IL-6 and TNFα), adhesion molecules (ICAM1, VCAM1, and MCP1), and extracellular matrix degrading enzymes (MMP1 and MMP9). The control exerted by *NEXN* on these molecular processes may also come into play in lung cancer. Here we showed that *NEXN* knockdown lung fibroblasts do not show altered endogenous DNA damage, implying the need of investigating alternative mechanisms of action in future functional studies.

15q25 is the locus most strongly associated with lung cancer[1], but also a leading susceptibility locus for smoking behavior[23] and other traits related to lung disease such as chronic obstructive pulmonary disease (COPD)[24]. COPD and lung cancer-associated variants in 15q25 are known expression and methylation QTL (eQTL and meQTL) for multiple genes and tissues[3, 25]. It has not been possible so far to definitely identify all of the causal gene(s) at this locus, but most evidence points toward *CHRNA5* (cholinergic receptor nicotinic alpha 5 subunit) or *IREB2* (iron responsive element binding protein 2). In this study, we focused specifically on gene expression in lung tissues with the hope to identify genes directly involved in lung cancer development. More than one Bonferroni-corrected TWAS genes were identified at 15q25. The top one was *IREB2*, and then in order of significance, *CHRNA3*, *CHRNA5*, *HYKK*, and *PSMA4*. *IREB2* was also the top significant TWAS gene at this locus for COPD[8], and the results is in line with previous analysis specifically focused on 15q25[26]. *IREB2* encodes a RNA-binding protein that regulates iron levels in cells. Alteration of iron metabolism has been observed in NSCLC[27] and iron has been shown to influence apoptosis of lung cancer cells (A549)[28]. Silencing of *IREB2* in these cells has been shown to modulate the expression of iron metabolism-related genes (transferrin receptor and ferritin)[29] and injection of wild-type *IREB2* in mice was shown to stimulate growth of tumor xenografts[30]. Previous studies have thus demonstrated a potential biological link between *IREB2* and lung cancer. In the current study, knockdown of *IREB2* was showed to increase endogenous DNA damage in human lung fibroblasts, supporting a potential cancer-promoting role in the lung by elevated DNA damage and genomic instability. However, the 15q25 locus harbors additional candidate genes including three nicotinic receptors, namely *CHRNA3*, *CHRNA5*, and *CHRNB4*. Variation in these genes have been strongly associated with smoking behavior and other aspects of addiction, thus indirectly affecting lung cancer risk through modulation of smoking behavior[31]. It should be emphasized that our study is relevant for

lung expression only and that causal genes of addiction to smoking on 15q25 may be complemented by future brain eQTL studies. Similarly, other forms of genetic variation may be modulating function at this locus, for example one most associated SNPs at this locus encodes a missense change in *CHRNA5* (rs16969968). Our results nevertheless suggest the possibility that one or more genes acting in the lung, brain or other tissues may mediate the risk of lung cancer on 15q25. The *IREB2* locus shows linkage disequilibrium with the *CHRNA3/CHRNA5/CHRNB4* locus complicating our ability to distinguish between these genetic effects. Clearly, more research will be needed to pinpoint the causal gene(s) or pathway(s) underpinning this lung cancer susceptibility locus.

On 6p22-p21 (MHC locus), multiple candidate causal genes were identified for overall lung cancer, squamous cell carcinoma and ever-smokers. However, no TWAS gene was found for adenocarcinoma, which is consistent with previous GWAS showing stronger association with squamous cell carcinoma at the MHC locus[6, 32]. The interpretation of TWAS results in the MHC locus is complicated by the extended LD structure in this region. TWAS cannot distinguish causal relationship and pleiotropy. For example, if the same SNPs affect the expression level of more than one gene, TWAS cannot delineate the causal one. Here, we identified multiple candidate genes on 6p22-p21 that varied by histological subtypes and that showed some similarity, but also differences between S-PrediXcan and FUSION. Although the top TWAS gene with both TWAS approaches was *APOM*, our study does not provide firm conclusion about the most likely causal gene(s) in the MHC locus and suggests the need of using alternative methods to reach this goal in this region.

It should be emphasized that TWAS genes do not imply causality. TWAS genes are more appropriately interpreted as prioritized or ranked candidate causal genes at loci[33]. In addition, TWAS cannot distinguish causal relationship and pleiotropy. For example, if the same SNPs affect the expression level of more than one gene, TWAS cannot delineate the causal one. In this study, we intentionally highlighted the top TWAS finding at each locus. It is not uncommon to observe multiple TWAS genes per locus, which is caused by co-regulation and shared eQTL[34]. Further functional experiments will be needed to demonstrate causality of one or more genes at each locus. One of the main limitation of TWAS is to study only genes with a significant cis-heritability, i.e. genes for which a part of expression can be explained by SNPs. This leaves out a large proportion of genes including known and potential cancer genes, particularly variants that influence gene product function through other ways. On the other hand, by focusing on the genetic component of expression, we avoid confounding effects of other factors (measured or not) on gene expression. This however does not preclude confounders of the SNP-expression correlation derived from the lung eQTL mapping study. We have used bulk gene expression data from the lung in both the discovery and validation (GTEx) sets. The lung is a heterogeneous tissue containing many cell types (organ-specific and migratory) with relative proportions that can vary based on the underlying lung disease, harvesting location, histological subtypes, and environmental factors[35]. These factors may have limited our ability to derive lung eQTL signals and subsequently study by TWAS the association between the cis-genetic component of expression and lung cancer. In addition, with our approach, we were unable to identify cell type-specific eQTL signals including from rare (or less frequent) cell types that may give rise to cancer that are not well represented in bulk expression data.

In conclusion, this work outlines a new lung adenocarcinoma locus on 9p13.3 with *AQP3* as the most likely underlying causal gene. Within known lung cancer GWAS loci, we map *IREB2* on 15q25 for all histological subtypes and ever-smokers, *NEXN* on 1p31.1 in never-smokers, and provide putative causal genes for 15 additional loci. The cancer-promoting role of *IREB2* and *AQP3* were further supported by endogenous DNA damage assays in human lung fibroblasts. TWAS genes are key to understand disease etiology, facilitate biological interpretation of GWAS results, and prioritize follow-up functional studies.

Acknowledgements

The authors would like to thank the team at the IUCPQ site of the Respiratory Health Network (RHN) Tissue Bank of the FRQS for their valuable assistance.

Financial support

Yohan Bossé holds a Canada Research Chair in Genomics of Heart and Lung Diseases. This study was supported by grants from the Chaire de pneumologie de la Fondation JD Bégin de l'Université Laval, the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec, and the Canadian Institutes of Health Research (MOP – 123369) to Yohan Bossé and Institut national du Cancer (France) to James McKay (TABAC 17-022). CARET is funded by the National Cancer Institute, National Institutes of Health through grants U01-CA063673, UM1-CA167462, and U01-CA167462. The genetic data was supported by NIH U19 CA148127 and U19 CA203654. The *in vitro* assays was supported by National Institutes of Health (NIH) Director's Pioneer Award DP1-CA174424 (SMR); the WM Keck Foundation (SMR); R35-GM122598 (SMR); the BCM Cytometry and Cell Sorting Core with funding from the NIH P30-AI036211, P30-CA125123, and S10-RR024574.

References

1. Bossé Y, Amos CI. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol Biomarkers Prev* 2018;**27**: 363-79.

2. Bossé Y. Genome-wide expression quantitative trait loci analysis in asthma. *Curr Opin Allergy Clin Immunol* 2013;**13**: 487-94.

3. Nguyen JD, Lamontagne M, Couture C, Conti M, Pare PD, Sin DD, Hogg JC, Nickle D, Postma DS, Timens W, Laviolette M, Bossé Y. Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. *Carcinogenesis* 2014;**35**: 2653-9.

4. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;**48**: 245-52.

5. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, Stahl EA, Huckins LM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 2018;**9**: 1825.

6. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017;**49**: 1126-32.

7. Hao K, Bossé Y, Nickle DC, Pare PD, Postma DS, Laviolette M, Sandford A, Hackett TL, Daley D, Hogg JC, Elliott WM, Couture C, et al. Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *PLoS Genet* 2012;**8**: e1003029.

8. Lamontagne M, Berube JC, Obeidat M, Cho MH, Hobbs BD, Sakornsakolpat P, de Jong K, Boezen HM, International CGC, Nickle D, Hao K, Timens W, et al. Leveraging lung

tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Hum Mol Genet* 2018;**27**: 1819-29.

9. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**: 118-27.

10. Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* 2019;**51**: 768-9.

11. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017;**550**: 204-13.

12. Xia J, Chiu LY, Nehring RB, Bravo Nunez MA, Mei Q, Perez M, Zhai Y, Fitzgerald DM, Pribis JP, Wang Y, Hu CW, Powell RT, et al. Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. *Cell* 2019;**176**: 127-43 e24.

13. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001;**25**: 402-8.

14. Liu YL, Matsuzaki T, Nakazawa T, Murata S, Nakamura N, Kondo T, Iwashina M, Mochizuki K, Yamane T, Takata K, Katoh R. Expression of aquaporin 3 (AQP3) in normal and neoplastic lung tissues. *Hum Pathol* 2007;**38**: 171-8.

15. Hanada S, Maeshima A, Matsuno Y, Ohta T, Ohki M, Yoshida T, Hayashi Y, Yoshizawa Y, Hirohashi S, Sakamoto M. Expression profile of early lung adenocarcinoma: identification of MRP3 as a molecular marker for early progression. *J Pathol* 2008;**216**: 75-82.

16. Xiong G, Chen X, Zhang Q, Fang Y, Chen W, Li C, Zhang J. RNA interference influenced the proliferation and invasion of XWLC-05 lung cancer cells through inhibiting aquaporin 3. *Biochem Biophys Res Commun* 2017;**485**: 627-34.

17. Hou SY, Li YP, Wang JH, Yang SL, Wang Y, Wang Y, Kuang Y. Aquaporin-3 Inhibition Reduces the Growth of NSCLC Cells Induced by Hypoxia. *Cell Physiol Biochem* 2016;**38**: 129-40.

18. Xia H, Ma YF, Yu CH, Li YJ, Tang J, Li JB, Zhao YN, Liu Y. Aquaporin 3 knockdown suppresses tumour growth and angiogenesis in experimental non-small cell lung cancer. *Exp Physiol* 2014;**99**: 974-84.

19. Hassel D, Dahme T, Erdmann J, Meder B, Huge A, Stoll M, Just S, Hess A, Ehlermann P, Weichenhan D, Grimmler M, Liptau H, et al. Nexilin mutations destabilize cardiac Z-disks and lead to dilated cardiomyopathy. *Nat Med* 2009;**15**: 1281-8.

20. Wang H, Li Z, Wang J, Sun K, Cui Q, Song L, Zou Y, Wang X, Liu X, Hui R, Fan Y. Mutations in NEXN, a Z-disc gene, are associated with hypertrophic cardiomyopathy. *Am J Hum Genet* 2010;**87**: 687-93.

21. Yuan H, Liu H, Liu Z, Owzar K, Han Y, Su L, Wei Y, Hung RJ, McLaughlin J, Brhane Y, Brennan P, Bickeboeller H, et al. A Novel Genetic Variant in Long Non-coding RNA Gene NEXN-AS1 is Associated with Risk of Lung Cancer. *Sci Rep* 2016;**6**: 34234.

22. Hu YW, Guo FX, Xu YJ, Li P, Lu ZF, McVey DG, Zheng L, Wang Q, Ye JH, Kang CM, Wu SG, Zhao JJ, et al. Long noncoding RNA NEXN-AS1 mitigates atherosclerosis by regulating the actin-binding protein NEXN. *J Clin Invest* 2019.

23. Tobacco Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010;**42**: 441-7.

24. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, Shrine N, Qiao D, Bartz TM, Kim DK, Lee MK, Latourelle JC, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet* 2019;**51**: 494-505.

25. Nedeljkovic I, Carnero-Montoro E, Lahousse L, van der Plaat DA, de Jong K, Vonk JM, van Diemen CC, Faiz A, van den Berge M, Obeidat M, Bosse Y, Nickle DC, et al. Understanding the role of the chromosome 15q25.1 in COPD through epigenetics and transcriptomics. *Eur J Hum Genet* 2018;**26**: 709-22.

26. Fehringer G, Liu G, Pintilie M, Sykes J, Cheng D, Liu N, Chen Z, Seymour L, Der SD, Shepherd FA, Tsao MS, Hung RJ. Association of the 15q25 and 5p15 lung cancer susceptibility regions with gene expression in lung tumor tissue. *Cancer Epidemiol Biomarkers Prev* 2012;**21**: 1097-104.

27. Kukulj S, Jaganjac M, Boranic M, Krizanac S, Santic Z, Poljak-Blazi M. Altered iron metabolism, inflammation, transferrin receptors, and ferritin expression in non-small-cell lung cancer. *Med Oncol* 2010;**27**: 268-77.

28. Choi SJ, Oh JM, Choy JH. Toxicological effects of inorganic nanoparticles on human lung cancer A549 cells. *J Inorg Biochem* 2009;**103**: 463-71.

29. Cheng Z, Dai LL, Song YN, Kang Y, Si JM, Xia J, Liu YF. Regulatory effect of iron regulatory protein-2 on iron metabolism in lung cancer. *Genet Mol Res* 2014;**13**: 5514-22.

30. Maffettone C, Chen G, Drozdov I, Ouzounis C, Pantopoulos K. Tumorigenic properties of iron regulatory protein 2 (IRP2) mediated by its specific 73-amino acids insert. *PLoS One* 2010;**5**: e10163.

31. Chen LS, Baker TB, Piper ME, Breslau N, Cannon DS, Doheny KF, Gogarten SM, Johnson EO, Saccone NL, Wang JC, Weiss RB, Goate AM, et al. Interplay of genetic risk factors (CHRNA5-CHRNA3-CHRNB4) and cessation treatments in smoking cessation success. *Am J Psychiatry* 2012;**169**: 735-42.

32. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, Risch A, McKay JD, Wang Y, Dai J, Gaborieau V, McLaughlin J, et al. Influence of common genetic

variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* 2012;**21**: 4980-95.

33. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, Bjorkegren JLM, Im HK, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* 2019;**51**: 592-9.

34. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, Song L, Safi A, Schizophrenia Working Group of the Psychiatric Genomics C, McCarroll S, Neale BM, Ophoff RA, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat Genet* 2018;**50**: 538-48.

35. McCall MN, Illei PB, Halushka MK. Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. *Am J Hum Genet* 2016;**99**: 624-35.

Table 1. TWAS genes identified by lung cancer and smoking subgroups

Subgroups	**Loci**	**Bonferroni-corrected TWAS genes[1]**		**GTEx**	
		S-PrediXcan (direction)	**FUSION (direction)**	**S-PrediXcan (direction, P_{TWAS})**	**FUSION (direction, P_{TWAS})**
Overall lung cancer	15q25	*IREB2* (-) > *CHRNA3* (-) > *CHRNA5* (-) > *HYKK* (-) > *PSMA4* (+)	*IREB2* (-) > *CHRNA5* (-) > *HYKK* (-) > *PSMA4* (+)	*IREB2* (no model) *CHRNA3* (-, 1.33E-12) *CHRNA5* (-, 1.70E-14) *HYKK* (no model) *PSMA4* (no model)	*IREB2* (no model) *CHRNA3* (-, 3.54E-23) *CHRNA5* (-, 1.71E-14) *HYKK* (no model) *PSMA4* (no model)
	12p13.33	*RAD52* (+)		*RAD52* (+, 2.10E-9)	*RAD52* (+, 1.43E-09)
	15q21.1	*CTSH* (+) > *SECISBP2L* (-)	*SECISBP2L* (-)	*CTSH* (+, 0.00015) *SECISBP2L* (-, 5.79E-8)	*CTSH* (+, 0.0001) *SECISBP2L* (-, 4.42E-08)
	6p-MHC	*APOM* (-) > *ZFP57* (+) > *HLA-A* (-) > *TRIM38* (-) > *HLA-F* (+) > *ZKSCAN3* (-) > *HLA-DPB1* (-) > *BAT1* (-) > *HLA-DPA1* (+)	*APOM* (-) > *ZNRD1* (-) > *PRRC2A* (-) > *ZFP57* (+) > *FLOT1* (+) > *NOTCH4* (-) > *TUBB* (+) > *HLA-A* (-) > *HLA-F* (+) > *HLA-G* (-) > *HCG8* (-) > *HLA-J* (-) > *TRIM38* (-) > *HCP5* (-) > *SFTA2* (+) > *HLA-L* (-) > *HLA-DQB1* (-) > *ZKSCAN3* (-) > *PSORS1C3* (-) > *CCHCR1* (+)	*CCHCR1* (+, 5.08E-16)[2]	*FLOT1* (+, 6.05E-16)[2]
	6q27	*RNASET2* (+)	*FGFR1OP* (-)	*RNASET2* (+, 1.16E-6) *FGFR1OP* (no model)	*RNASET2* (+, 1.33E-07) *FGFR1OP* (-, 6.23E-04)
	11q23.3	*AMICA1/JAML* (-)	*JAML* (-)	*JAML* (no model)	*JAML* (no model)
Adenocarcinoma	15q25	*IREB2* (-) > *CHRNA3* (-) > *CHRNA5* (-) > *HYKK* (-)	*IREB2* (-) > *CHRNA5* (-) > *HYKK* (-) > *PSMA4* (+)	*IREB2* (no model) *CHRNA3* (-, 1.96E-07) *CHRNA5* (-, 1.23E-8) *HYKK* (no model) *PSMA4* (no model)	*IREB2* (no model) *CHRNA3* (-, 1.87E-12) *CHRNA5* (-, 7.16E-09) *HYKK* (no model) *PSMA4* (no model)
	15q21.1	*SECISBP2L* (-) > *GALK2* (+)	*SECISBP2L* (-) > *GALK2* (+) > *FAM227B* (-)	*SECISBP2L* (-, 2.57E-14) *GALK2* (no model) *FAM227B* (+, 1.13E-7)	*SECISBP2L* (-, 1.54E-14) *GALK2* (no model) *FAM227B* (+, 8.53E-05)
	3q28	*TP63* (-)	*TP63* (-)	*TP63* (-, 3.64E-6)	*TP63* (no model)
	11q23.3	*JAML* (-)	*JAML* (-)	*JAML* (-, 0.158)	*JAML* (no model)
	8p12	*NRG1* (-)	*NRG1* (-)	*NRG1* (-, 2.67E-6)	*NRG1* (no model)
	9p13.3	***AQP3*** (+, P_{TWAS}=3.72E-6)	***AQP3*** (+)	*AQP3* (+, 6.55E-5)	*AQP3* (+, 1.72E-05)
	6q22.1	*DCBLD1* (-)		*DCBLD1* (-, 8.69E-7)	*DCBLD1* (-, 4.63E-07)
Squamous cell	15q25	*IREB2* (-)> *CHRNA3* (-) > *CHRNA5* (-)	*IREB2* (-)	*IREB2* (no model) *CHRNA3* (-, 0.00095) *CHRNA5* (-, 0.00063)	*IREB2* (no model) *CHRNA3* (-, 4.40E-07) *CHRNA5* (-, 6.78E-04)
	6p-MHC	*APOM* (-) > *HLA-DQB1* (-) > *TRIM38* (-) > *ZKSCAN3* (-) >	*APOM* (-) > *NOTCH4* (-) > *PRRC2A* (-) > *HCP5* (-) >	*ATF6B* (+, 3.70E-15)[2]	*C4A* (-, 7.52E-15)[2]

		HLA-DPB1 (-) > *HLA-DQB2* (-) > *ZNF389* (-) > *ZNF187* (-) *HLA-A* (-) > ZFP57 (+) > HIST1H2AA (-) > LST1 (+)	*HLA-DQB1* (-) > *FLOT1* (+) > *ZNRD1* (-) > *MICB* (+) > *TRIM38* (-) > *TUBB* (+) > *HLA-A* (-) > *ZKSCAN3* (-) > *HCG8* (-) > *ZNF192P1* (-) > *ZSCAN26* (-) > *HLA-L* (-) > *ZFP57* (+) > *HLA-F* (+) > *HCG14* (+)		
	10q24.31		***BLOC1S2*** (-)	*BLOC1S2* (-, 4.99E-7)	*BLOC1S2* (-, 8.36E-07)
	12p13.33	*RAD52* (+)		*RAD52* (+, 1.22E-10)	*RAD52* (+, 5.53E-12)
SCLC	15q25	*IREB2* (-) > CHRNA3 (-)	*IREB2* (-)	*IREB2* (no model) *CHRNA3* (-, 0.023)	*IREB2* (no model) *CHRNA3* (-, 3.51E-05)
	6p-MHC		*HIST1H2BD* (+)	*HIST1H2BD* (no model)	*HIST1H2BD* (no model)
	4q32.2		***TMA16*** (+, P_{TWAS}=4.20E-6)	*TMA16* (no model)	*TMA16* (-, 9.38E-04)
Ever-smokers	15q25	*IREB2* (-) > *CHRNA3* (-) > *CHRNA5* (-) > *HYKK* (-) > *PSMA4* (+)	*IREB2* (-) > *CHRNA5* (-) > *HYKK* (-) > *PSMA4* (+)	*IREB2* (no model) *CHRNA3* (-, 9.19E-12) *CHRNA5* (-, 1.36E-13) *HYKK* (no model) *PSMA4* (no model)	*IREB2* (no model) *CHRNA3* (-, 3.47E-20) *CHRNA5* (-, 4.55E-13) *HYKK* (no model) *PSMA4* (no model)
	6p-MHC	*APOM* (-) > *ZFP57* (+) > *TRIM38* (-) > *BAT1* (-) > *HLA-A* (-)	*APOM* (-) > *ZNRD1* (-) > *PRRC2A* (-) > *ZFP57* (+) > *NOTCH4* (-) > *HLA-G* (-) > *TUBB* (+) > *HLA-F* (+) > *HCP5* (-) > *HLA-J* (-) > *HLA-A* (-) > *FLOT1* (+) > *HLA-DQB1* (-) > *CCHCR1* (+)	*CCHCR1* (+, 5.51E-15)[2]	*FLOT1* (+, 1.38E-14)[2]
	15q21.1	*SECISBP2L* (-)	*SECISBP2L* (-)	*SECISBP2L* (-, 8.73E-7)	*SECISBP2L* (-, 9.70E-7)
	12p13.33	*RAD52* (+)		*RAD52* (+, 1.06E-9)	*RAD52* (+, 2.23E-09)
Never-smokers	1p31.1	***NEXN*** (-, P_{TWAS}=3.11E-6)	***NEXN*** (-, P_{TWAS}=2.64E-5)	*NEXN* (-, 0.0059)	*NEXN* (-, 0.0028)

Bold indicates new loci or new susceptibility genes. Novel loci are defined as not overlapping (±500 Kb) with a previously reported GWAS lung cancer locus[1].

(+) and (-) indicate predicted gene expression positively or negatively associated with lung cancer risk.

[1]Specific P_{TWAS} values are provided for genes that did not pass the Bonferroni significance threshold.

[2]Only the top TWAS gene is indicated for the MHC locus.

Table 2. Top TWAS genes in GWAS-nominated loci for overall lung cancer

		Top TWAS genes (direction, P_{TWAS})		**Replication in GTEx (direction, P_{TWAS})**	
GWAS loci	**Suspected causal genes by GWAS alone***	**S-PrediXcan**	**FUSION**	**S-PrediXcan**	**FUSION**
1p36.32	*AJAP1*, *NPHP4*	*NPHP4* (+, 0.146)	*NPHP4* (+, 0.102)	*NPHP4* (+, 0.379)	*NPHP4* (+, 0.408)
1p31.1	*FUBP*, *DNAJB4*	*GIPC2* (+, 0.209)	*PIGK* (-, 0.037)	*GIPC2* (+, 0.187) *PIGK* (-, 0.176)	*GIPC2* (+, 0.194) *PIGK* (-, 0.2)
1q22	*MUC1*, *ADAM15*, *THBS3*	*THBS3* (+, 1.15E-05)	*DCST2* (+, 0.000172)	*THBS3* (+, 1.69E-09)	*THBS3* (+, 2.76E-05)
2p16.3	*NRXN1*				
2q32	*NUP35*	*DUSP19* (-, 0.097)	*NUP35* (+, 0.2349)	*DUSP19* (-, 0.702) *NUP35* (-, 0.458)	*DUSP19* (no model) *NUP35* (-, 0.341)
2q32.2	*HIBCH*, *INPP1*, *PMS1*, *STAT1*	***ORMDL1* (-, 0.029)**	***ORMDL1* (-, 0.028)**	*ORMDL1* (-, 0.082)	*ORMDL1* (-, 0.058)
3p26	No genes. Deletions associated with cancer	*SUMF1* (+, 0.083)	*SUMF1* (+, 0.0918)	*SUMF1* (+, 0.095)	*SUMF1* (+, 0.122)
3q28	*TP63*	*TP63* (-, 5.67E-06)	*TP63* (-, 1.99E-5)	*TP63* (-, 0.005)	*TP63* (no model)
3q29	*C3orf21*	*CPN2 (+, 0.122)*		No model	No model
4p15.2	*KCNIP4*	*KCNIP4* (-, 0.137)	*PACRGL* (+, 0.057)	*KCNIP4* (-, 0.293) *PACRGL* (+, 0.009)	*KCNIP4* (no model) *PACRGL* (+, 0.0193)
5p15	*TERT*, *CLPTM1L*	*SLC6A19 (+, 7.99E-06)*	*SLC6A3* (+, 0.0019)	*SLC6A19* (no model) *SLC6A3* (+, 0.00068)	*SLC6A19* (no model) *SLC6A3* (+, 0.00045)
5q14.2	*XRCC4*	*XRCC4* (+, 0.756)	*XRCC4* (+, 0.59)	No model	No model
5q31	*PAHA2*, *CSF2*, *IL3*, *SLC22A5*, *ACSL6*	***SLC22A5* (-, 0.0047)**	***SLC22A5* (-, 0.005)**	*SLC22A5* (-, 0.025)	*SLC22A5* (-, 0.018)
5q32	*STK32A*, *PPP2R2B*, *DPYSL3*	*STK32A* (+, 0.069)	*SPINK1* (-, 0.238)	*STK32A* (+, 0.569) *SPINK1* (-, 0.524)	*STK32A* (+, 0.424) *SPINK1* (-, 0.498)
6p22.2	*HIST1H1E*	***TRIM38* (-, 5.07E-09)**	***TRIM38* (-, 5.32E-08)**	No model	No model
6p21	*BAG6*, *APOM*, *TNXB*, *MSH5*, *BTNL2*, *PRRC2A*, *FKBPL*, *HSPA1B*, *FOXP4*, *FOXP4-AS1*, *GTF2H4*, *LRFN2*, *HLA-A*, *HLA-DQB1*	*APOM* (-, 3.16E-14)	*APOM* (-, 9.29E-16)	*APOM* (-, 0.0011)	*APOM* (no model)
6q22	*DCBLD1*, *ROS1*	*DCBLD1* (-, 0.0019)	*DCBLD1* (-, 0.0019)	*DCBLD1* (-, 0.0109)	*DCBLD1* (-, 0.00352)
6q27	*RNASET2*	*RNASET2* (+, 2.33E-08)	*FGFR1OP* (-, 7.68E-08)	*RNASET2* (+, 1.16E-06) *FGFR1OP* (no model)	*RNASET2* (+, 1.33E-07) *FGFR1OP* (-, 6.23E-04)
7p15.3	*SP4*, *DNAH11*	*FAM126A (-, 0.207)*		No model	No model
8p21.1	*EPHX2*, *CHRNA2*	*CLU (+, 1.78E-05)*	*CLU* (-, 0.00672)	*CLU* (no model)	No model

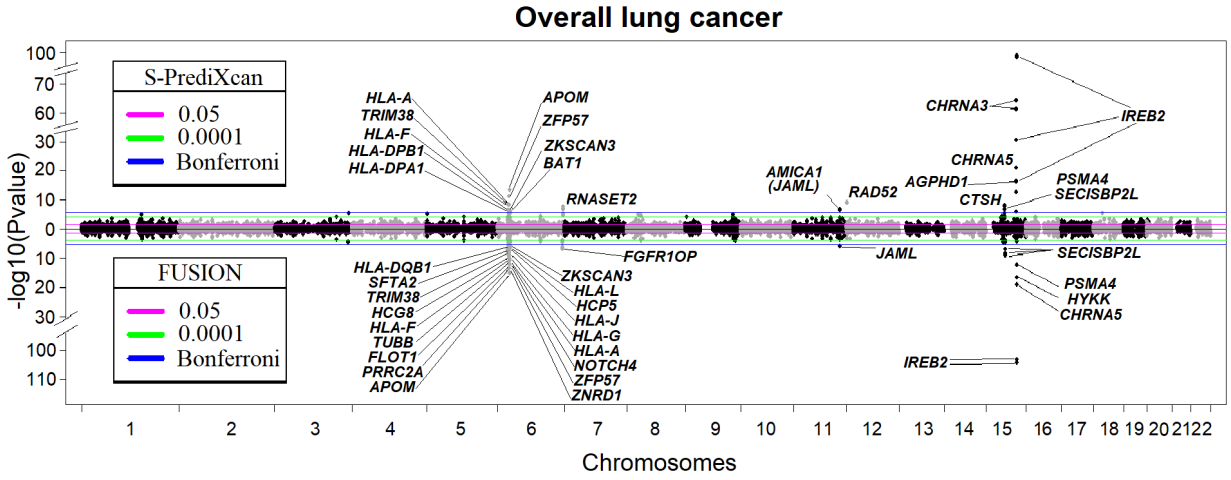
8p12	*NRG1*	*NRG1* (-, 3.37E-05)	*NRG1* (-, 9.69E-05)	*NRG1* (-, 0.0003)	No model
9p21.3	*CDKN2A*, *CDKN2B*, *CDKN2B-AS1*, *MTAP*	***MTAP* (-, 0.0013)**	***MTAP* (-, 0.0271)**	No model	*MTAP* (-, 0.0178)
10p14	*GATA3*	*GATA3 (-,0.633)*		No model	
10q23.33	*FFAR4*	*HECTD2* (+, 0.026)	*PPP1R3C* (+, 0.364)	No model	No model
10q24.3	*OBFC1*	*LZTS2* (-, 0.056)	*TMEM180* (+, 0.24)	*LZTS2* (No model) *TMEM180* (+, 0.193)	*LZTS2* (No model) *TMEM180* (+, 0.183)
10q25.2	*VTI1A*	*VTI1A* (-, 0469)	*VTI1A* (-, 0.34)	No model	No model
11q23.3	*MPZL3*, *JAML* (also known as *AMICA1*)	*JAML* (-, 2.64E-07)	*JAML* (-, 1.39E-06)	*JAML* (-, 0.135)	No model
12p13.33	*RAD52*	*RAD52* (+, 6.58E-10)	*RAD52* (+, 0.000363)	*RAD52* (+, 2.10E-09)	*RAD52* (+, 1.43E-09)
12q13.13	*ACVR1B*, *NR4A1*	*ACVR1B* (+, 0.0199)	*SLC11A2* (+, 0.0297)	*ACVR1B* (+, 0.039) *SLC11A2* (no model)	*ACVR1B* (+, 0.186) *SLC11A2* (no model)
12q23.1	*NR1H4*, *SLC17A8*	*ARL1* (+, 0.041)	*GOLGA2P5* (+, 0.036)	*ARL1* (+, 0.592) *GOLGA2P5* (no model)	*ARL1* (+, 0.155) *GOLGA2P5* (no model)
12q24	*SH2B3*	*TMEM116* (+, 0.006)	*ATXN2* (2, 0.019)	*TMEM116* (-, 0.234) *ATXN2* (No model)	*TMEM116* (-,0.348) *ATXN2* (No model)
13q12.12	*MIPEP*, *TNFRSF19*	*MIPEP* (-, 0.169)	*MIPEP* (-, 0.089)	*MIPEP* (-, 0.205)	*MIPEP* (-, 0.205)
13q13.1	*BRCA2*	***N4BP2L2* (-, 0.049)**	***N4BP2L2* (+, 0.027)**	No model	No model
13q31.3	*GPC5*	*GPC5 (-,0.387)*	*MIR17HG* (-, 0.947)	*GPC5 (-, 0.173) MIR17HG* (No model)	*GPC5 (-,0.211) MIR17HG* (No model)
15q21.1	*SEMA6D*, *SECISBP2L*	*SECISBP2L* (-, 5.44E-09)	*SECISBP2L* (-, 8.01E-10)	*SECISBP2L* (-, 5.79E-8)	*SECISBP2L* (-, 4.42E-8)
15q25	*CHRNA5*, *CHRNA3*, *CHRNB4*, *IREB2*, *PSMA4*, *HYKK*	*IREB2* (-, 1.09E-99)	*IREB2* (-, 4.97E-104)	No model	No model
17q24.3	*BPTF*	*BPTF* (-, 0.019)	*BPTF* (-, 0.016)	No model	No model
18p11.22	*FAM38B* (also known as *FAM38B2*), *APCDD1*, *NAPG*	*FAM38B2* (-, 0.167)	*APCDD1* (-, 0.318)	No model	No model
18q12.1	*GAREM*	*GALNT1* (-, 0.187)	*GAREM* (-, 0.00045)	No model	No model
19q13.2	*TGFB1*, *CYP2A6*	*ZNF565* (-, 0.005)	*C19orf54* (-, 0.005)	*ZNF565* (-, 0.02) *C19orf54* (-, 3.44E-5)	*ZNF565* (-, 0.006) *C19orf54* (-, 2.89E-05)
20q11.21	*BPIFB1*	*CPNE1* (-, 0.051)	*PIGU* (-, 0.027)	*CPNE1* (-, 0.453) *PIGU* (No model)	*CPNE1* (-, 0.41) *PIGU* (No model)
20q13.2	*CYP24A1*	*CBLN4* (-, 0.119)	*CBLN4* (-, 0.09)	No model	No model
20q13.33	*RTEL1*	*RTEL1 (+, 0.007)*		No model	No model
22q12.1	*CHEK2*	*PIK3IP1* (+, 0.009)	*XBP1* (+, 0.085)	*PIK3IP1* (No model)	*PIK3IP1* No model)

				XBP1 (+, 0.107)	*XBP1* (+, 0.047)
22q12.2	*LIF*, *HORMAD2*, *MTMR3*	***MTMR3* (+, 0.026)**	***MTMR3* (+, 0.025)**	*MTMR3* (+, 0.027)	*MTMR3* (+, 0.031)

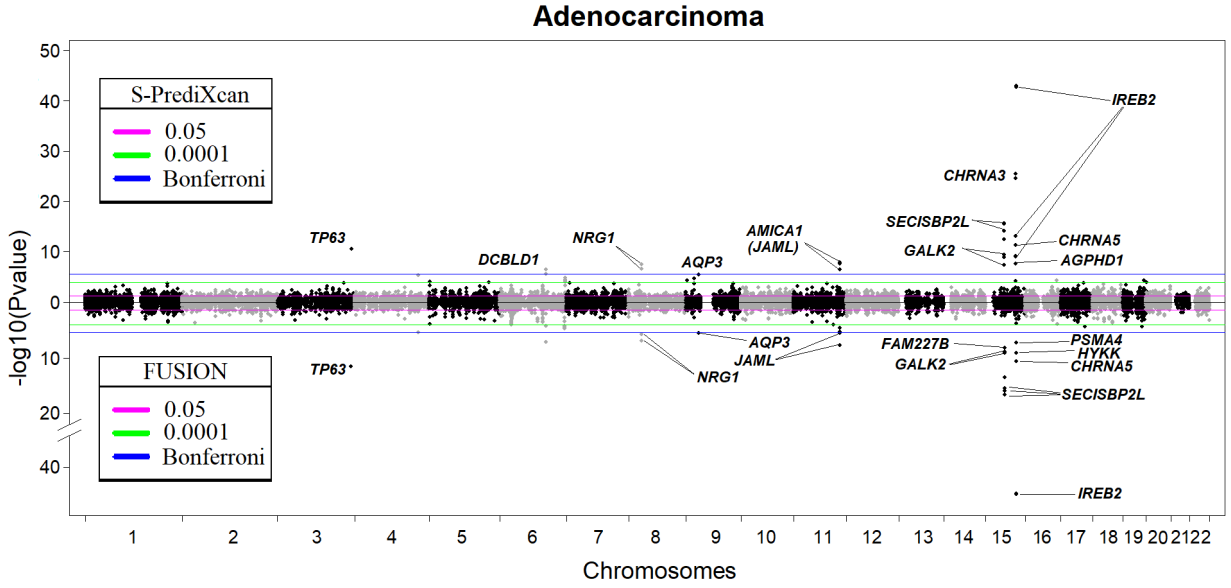
In bold are genes not in Table 1 showing some evidence of association ($P_{TWAS}<0.05$) with both S-PrediXcan and FUSION.
(+) and (-) indicate predicted gene expression positively or negatively associated with lung cancer risk.
*References are provided in Bossé and Amos[1].

Figures

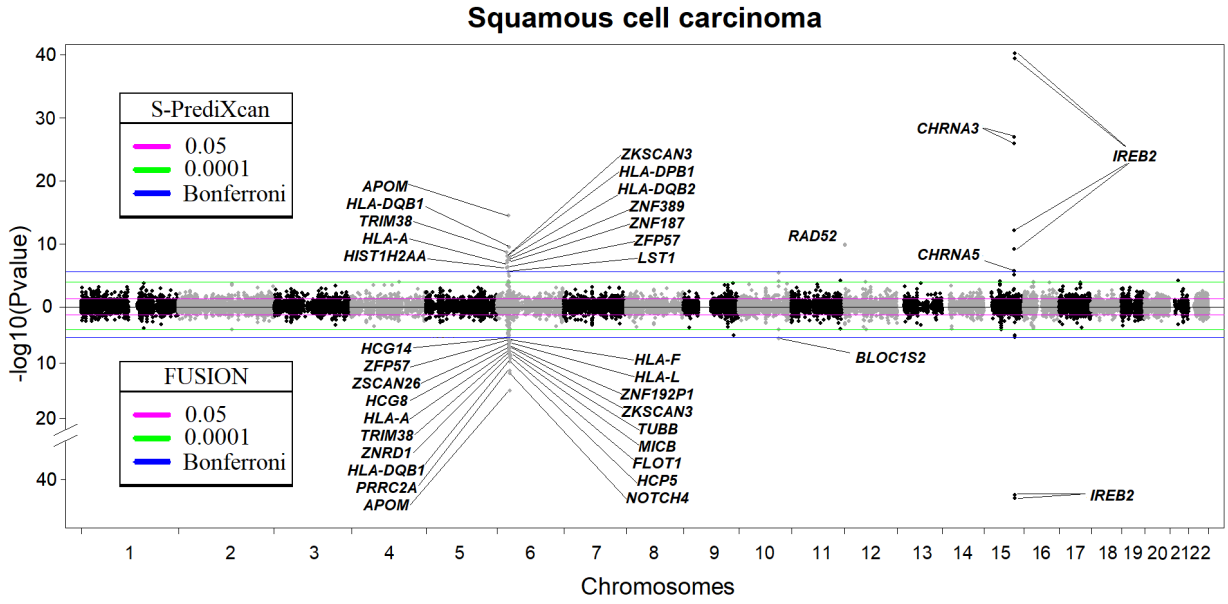
Figure 1. TWAS results for lung cancer overall, histological subtypes and smoking subgroups. Manhattan plots for S-PrediXcan (top) and FUSION (bottom) are illustrated in a mirror view to show similarity and differences between the two TWAS approaches. Each point represents a probe set with physical position plotted on the x-axis. The P values for gene expression-lung cancer associations are on the y-axis in -log10 scale. Annotations for the significant probe sets are indicated.
A) Overall lung cancer.



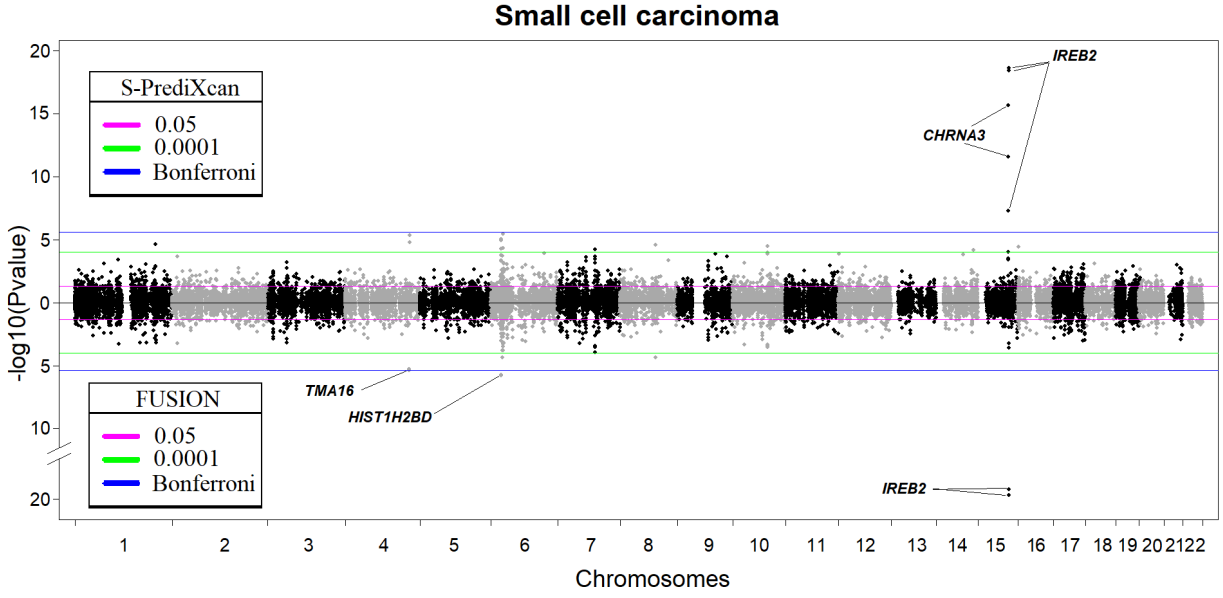
B) Adenocarcinoma.



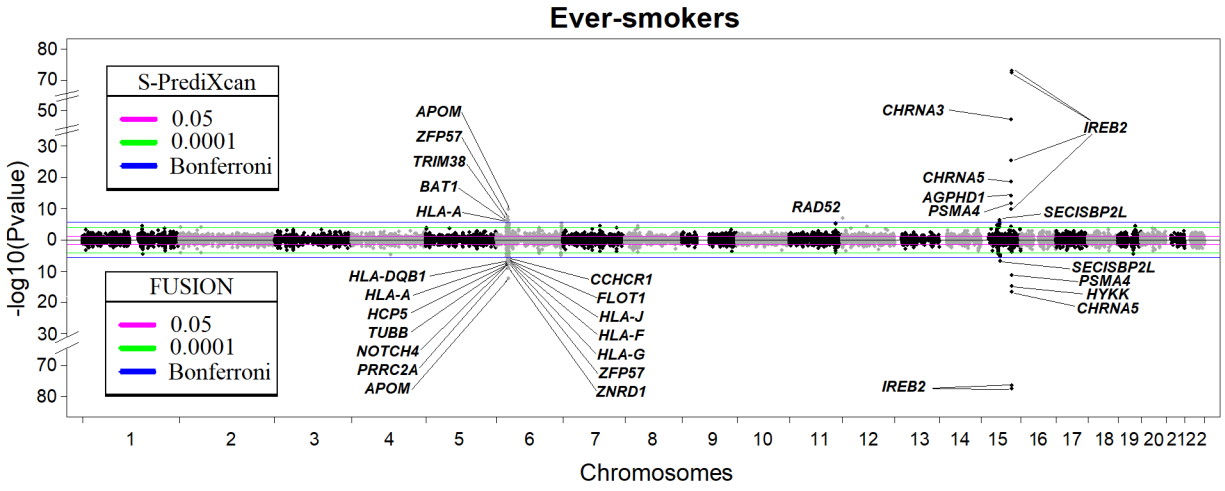
C) Squamous cell carcinoma.



D) Small cell lung cancer.



E) Ever-smokers.



F) Never-smokers.

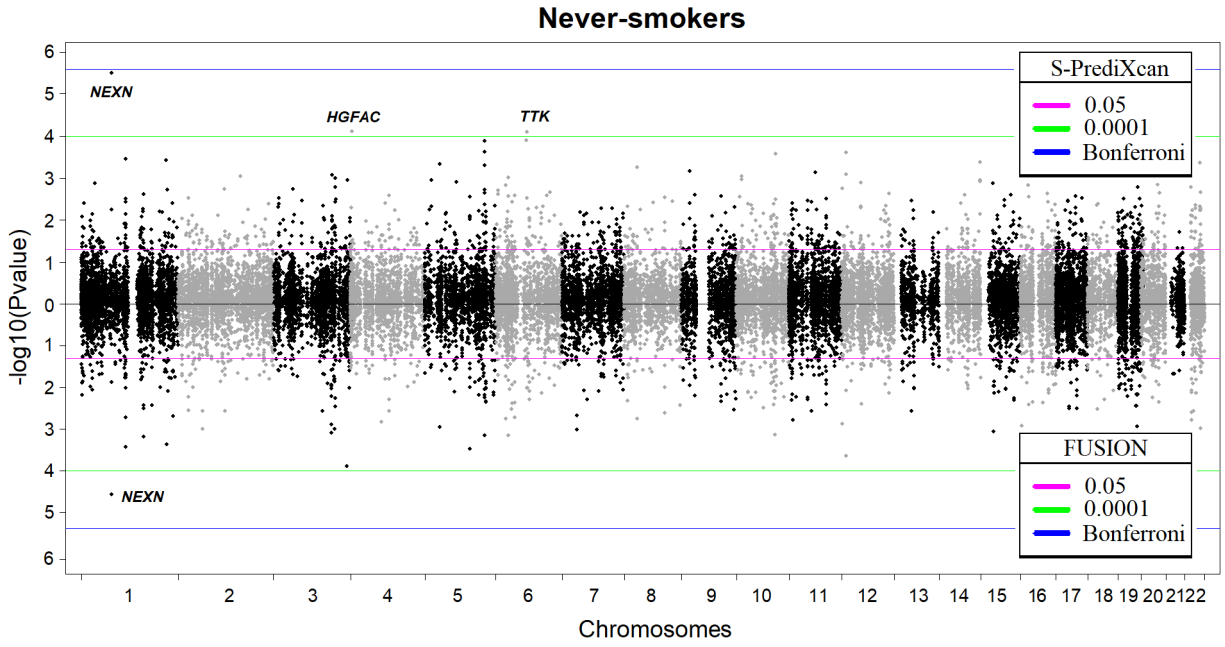


Figure 2. *IREB2* is the top candidate target gene on 15q25. The upper left panel shows the genetic associations with overall lung cancer in TRICL-ILCCO OncoArray. The bottom left panel shows the lung eQTL statistics for *IREB2*. The location of genes is illustrated at the bottom. The right panel shows boxplots of gene expression levels in the lung according to genotype groups for Laval, Groningen, and UBC samples. The y-axis shows the mRNA expression levels. The x-axis shows the three genotype groups for the SNP most strongly associated with lung cancer (upper right) and the SNP most strongly associated with mRNA expression of *IREB2* (lower right) with the number of individuals in parenthesis. The risk allele in TRICL-ILCCO OncoArray is shown in red.

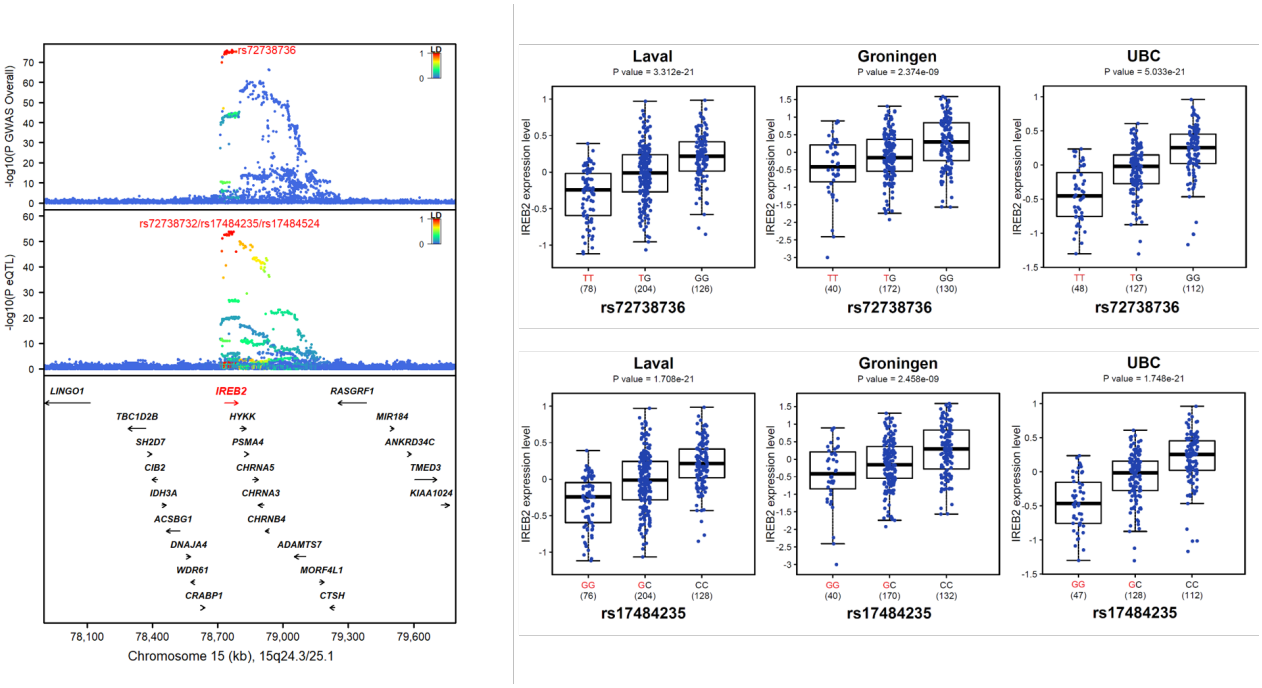


Figure 3. A novel susceptibility locus for adenocarcinoma on 9p13.3 with *AQP3* as the underlying gene. The upper left panel shows the genetic associations with adenocarcinoma in TRICL-ILCCO OncoArray. The bottom left panel shows the lung eQTL statistics for *AQP3*. The location of genes is illustrated at the bottom. The right panel shows boxplots of gene expression levels in the lung according to genotype groups for Laval, Groningen, and UBC samples. The y-axis shows the mRNA expression levels. The x-axis shows the three genotype groups for the SNP most strongly associated with lung cancer (upper right) and the SNP most strongly associated with mRNA expression of *AQP3* (lower right) with the number of individuals in parenthesis. The risk allele in TRICL-ILCCO OncoArray is shown in red.

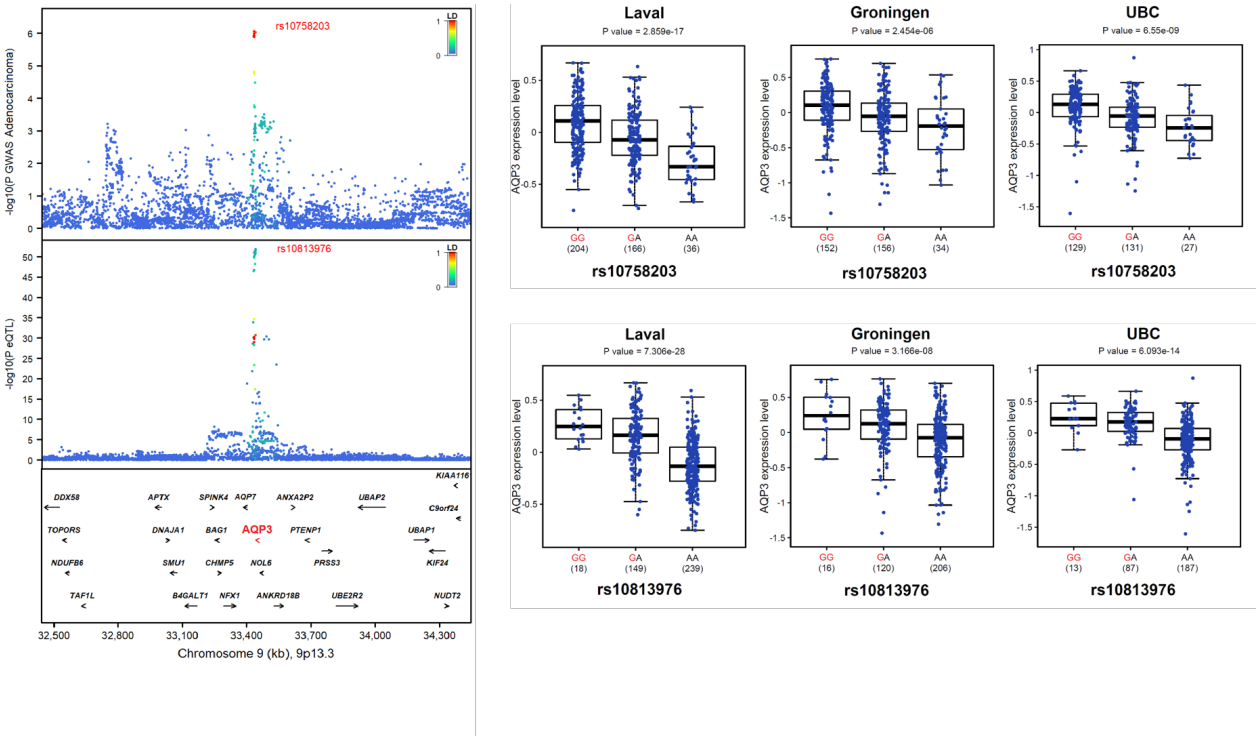


Figure 4. *NEXN* is the candidate target gene underpinning the lung cancer susceptibility locus for never-smokers on 1p31.1. The upper left panel shows the genetic associations with lung cancer in never-smokers in TRICL-ILCCO OncoArray. The bottom left panel shows the lung eQTL statistics for *NEXN*. The location of genes is illustrated at the bottom. The right panel shows boxplots of gene expression levels in the lung according to genotype groups for Laval, Groningen, and UBC samples. The y-axis shows the mRNA expression levels. The x-axis shows the three genotype groups for the SNP most strongly associated with lung cancer (upper right) and the SNP most strongly associated with mRNA expression of *NEXN* (lower right) with the number of individuals in parenthesis. The risk allele in TRICL-ILCCO OncoArray is shown in red.

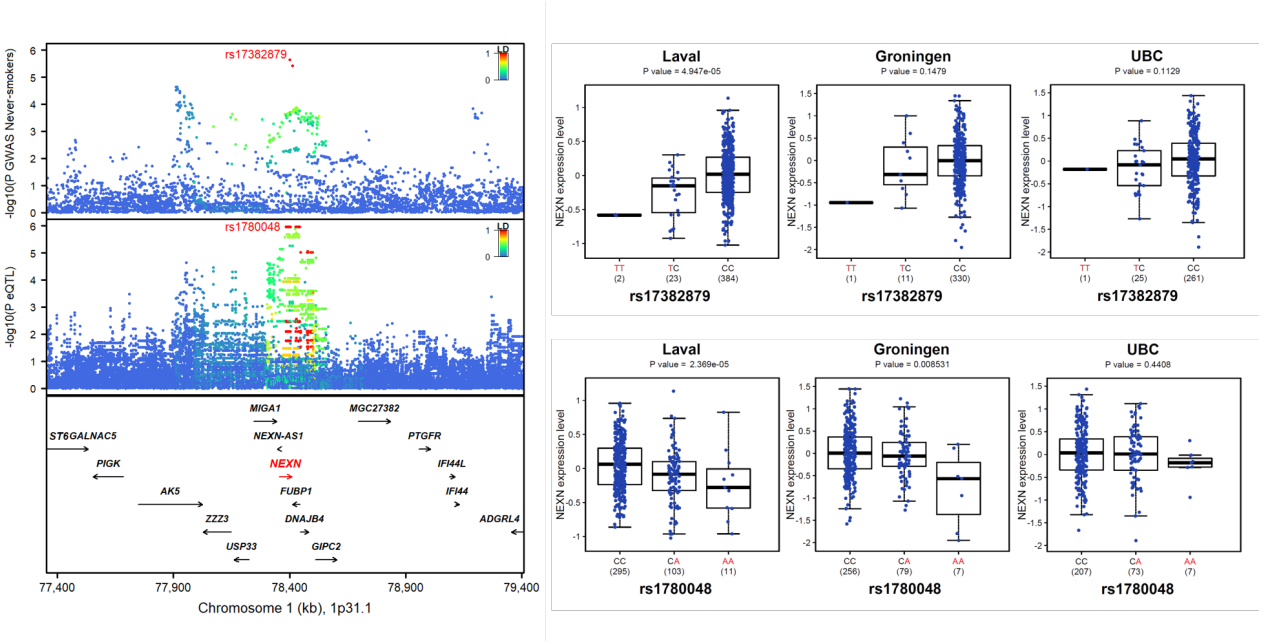
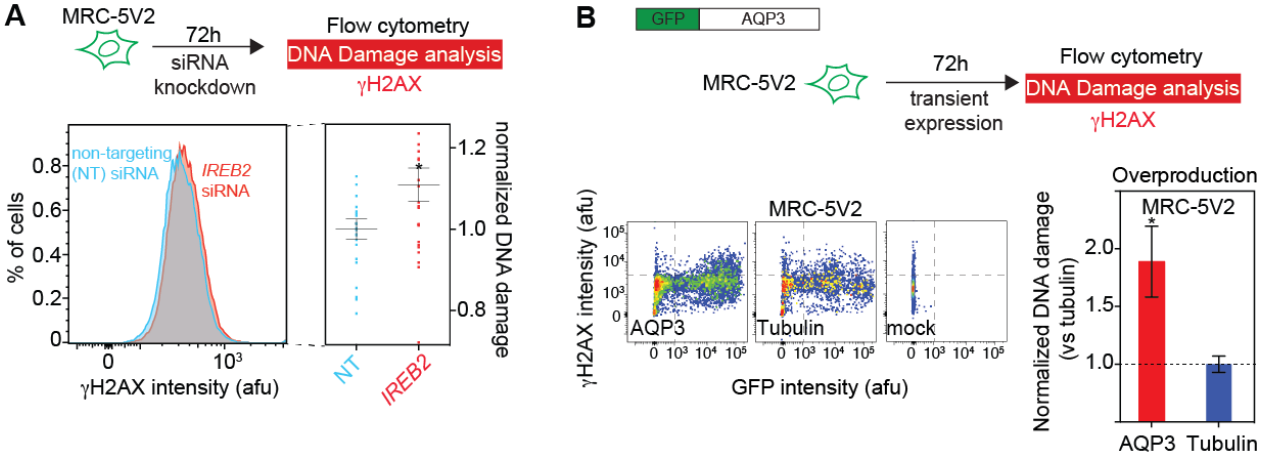
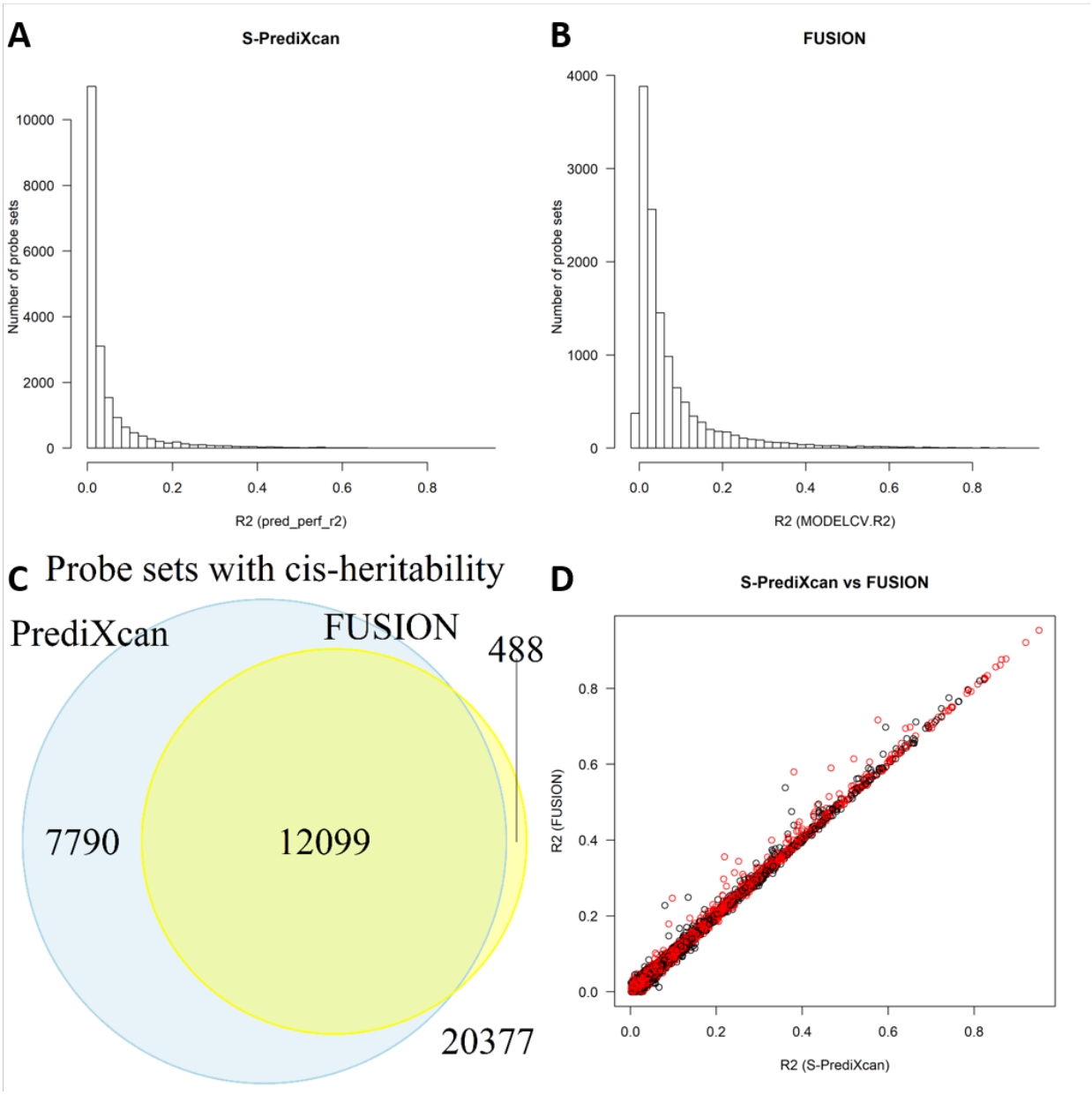


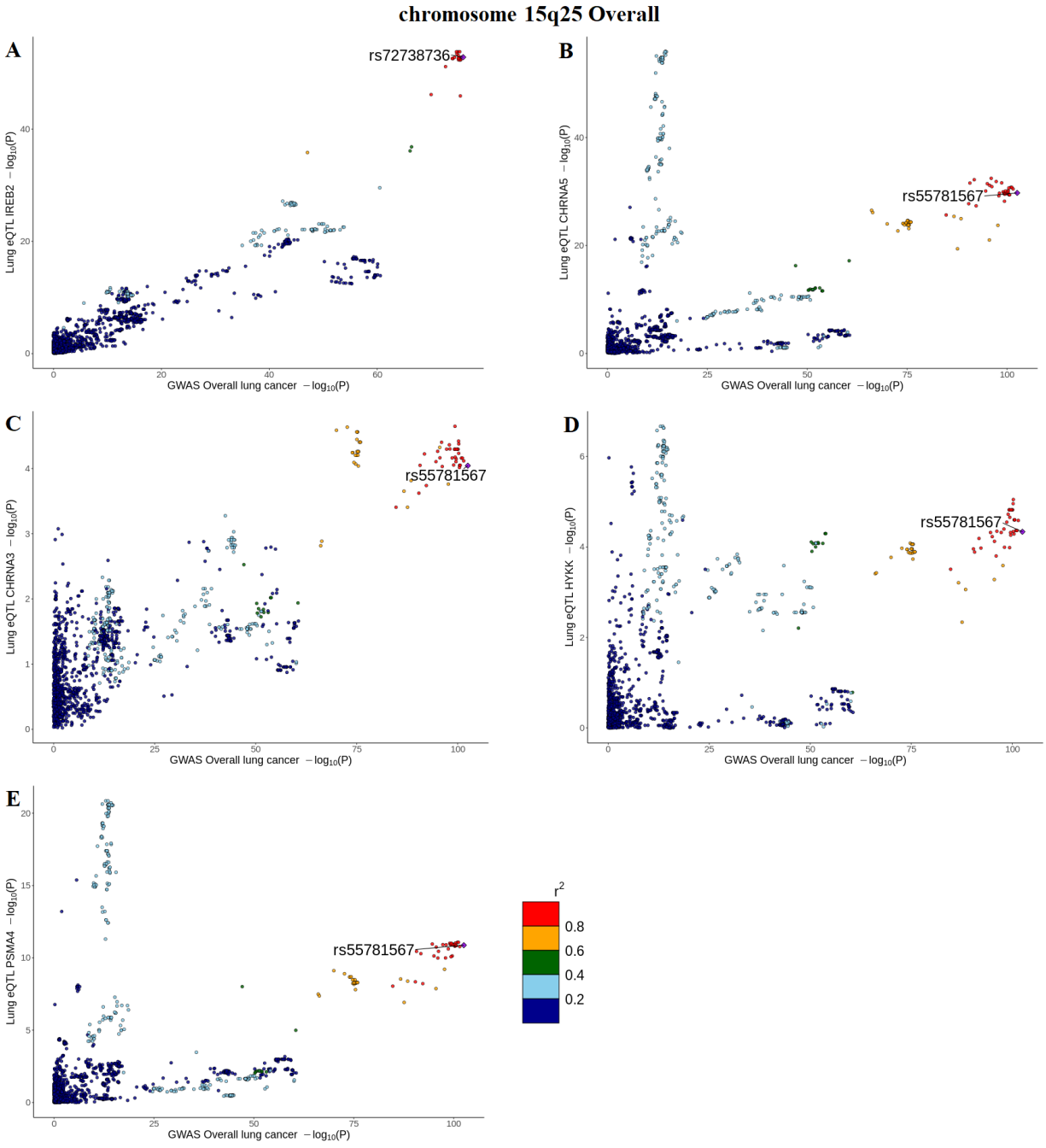
Figure 5. *IREB2* knockdown and AQP3 overproduction promote endogenous DNA damage. (A) Knocking down *IREB2*, but not *NEXN* by siRNA causes increased γH2AX level in MRC-5V2 cell line. Upper: scheme for a siRNA DNA damage assay. Lower left, representative flow cytometric histogram showing increased γH2AX in cells with *IREB2* knockdown compared with cells that were transfected with non-targeting siRNA. Lower right, a summary of at least three independent experiments for both *IREB2* and *NEXN*, n>20. Mean fluorescence intensity of each knockdown experiment was normalized to its corresponding non-targeting siRNA control. Error bar: SEM. (B) AQP3 overproduction increases endogenous γH2AX levels in MRC-5V2. Upper: full-length sequence-verified *AQP3* fused with N-terminal GFP fusion (and GFP-Tubulin as a control) were transiently overproduced in MRC-5V2 cell line and the DNA damage levels for both non-green and green cells were quantified by flow cytometry (details see methods). Lower left: representative flow cytometric histograms of GFP-*AQP3*, GFP-Tubulin, and mock transfection. Lower right: summary, mean ± SEM, n ≥8.



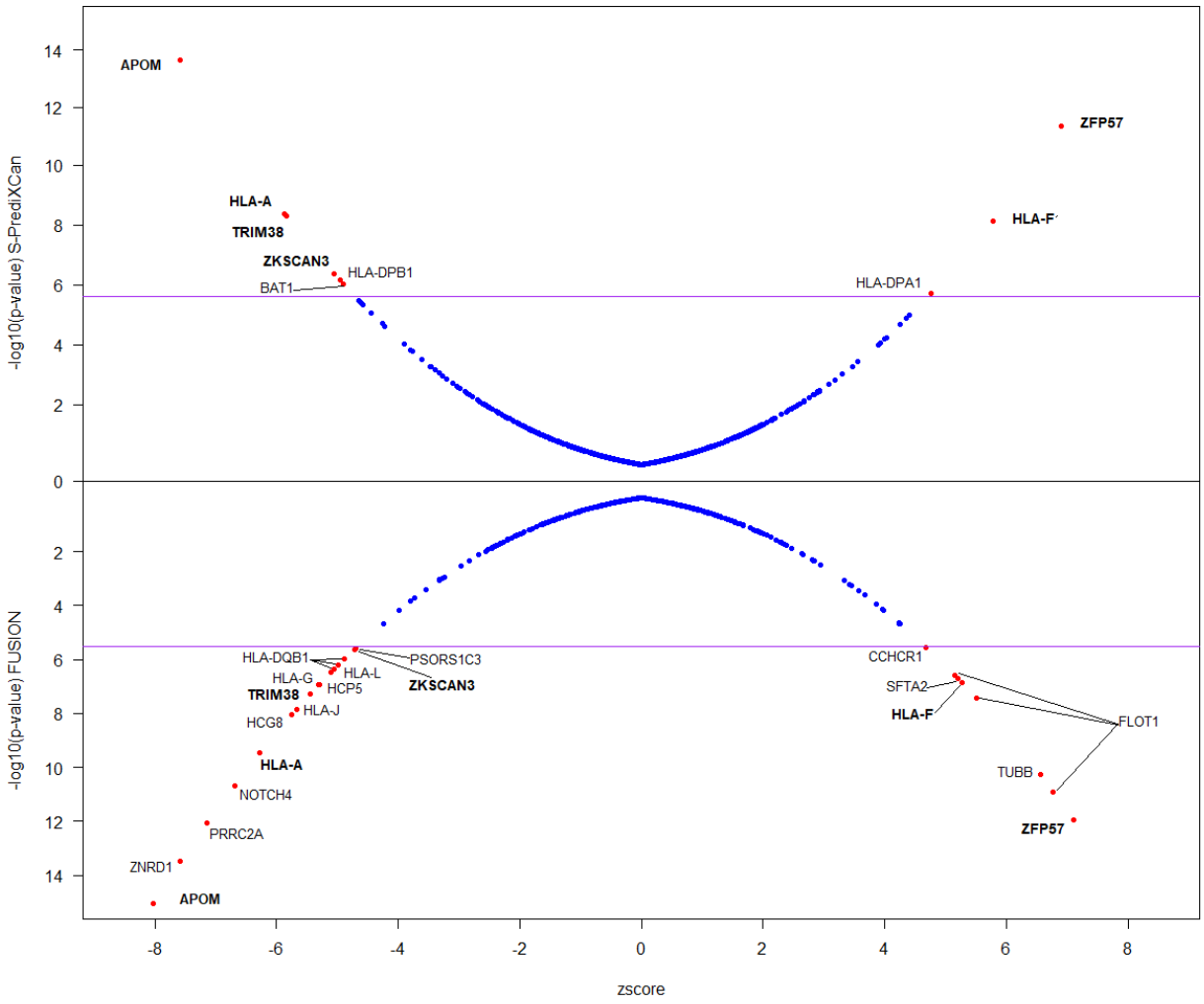
Supplementary Figure 1. Comparison of the *cis*-genetic component of expression in the lung between S-PrediXcan and FUSION. A) Histogram showing the distribution of the gene expression variance explained by SNPs for probe sets with significant prediction models (FDR<0.05) in S-PrediXcan. B) Histogram showing the distribution of the gene expression variance explained by SNPs for probe sets with significant *cis*-heritability (P<0.01) in FUSION. C) Venn diagram showing the number of probe sets with significant *cis*-genetic component of expression that overlap between S-PrediXcan and FUSION. D) Scatter plot showing the prediction performance for the 12,099 probe sets in common between S-PrediXcan and FUSION. Red dots indicate probe sets evaluated with different prediction models in S-PrediXcan and FUSION (enet vs LASSO).



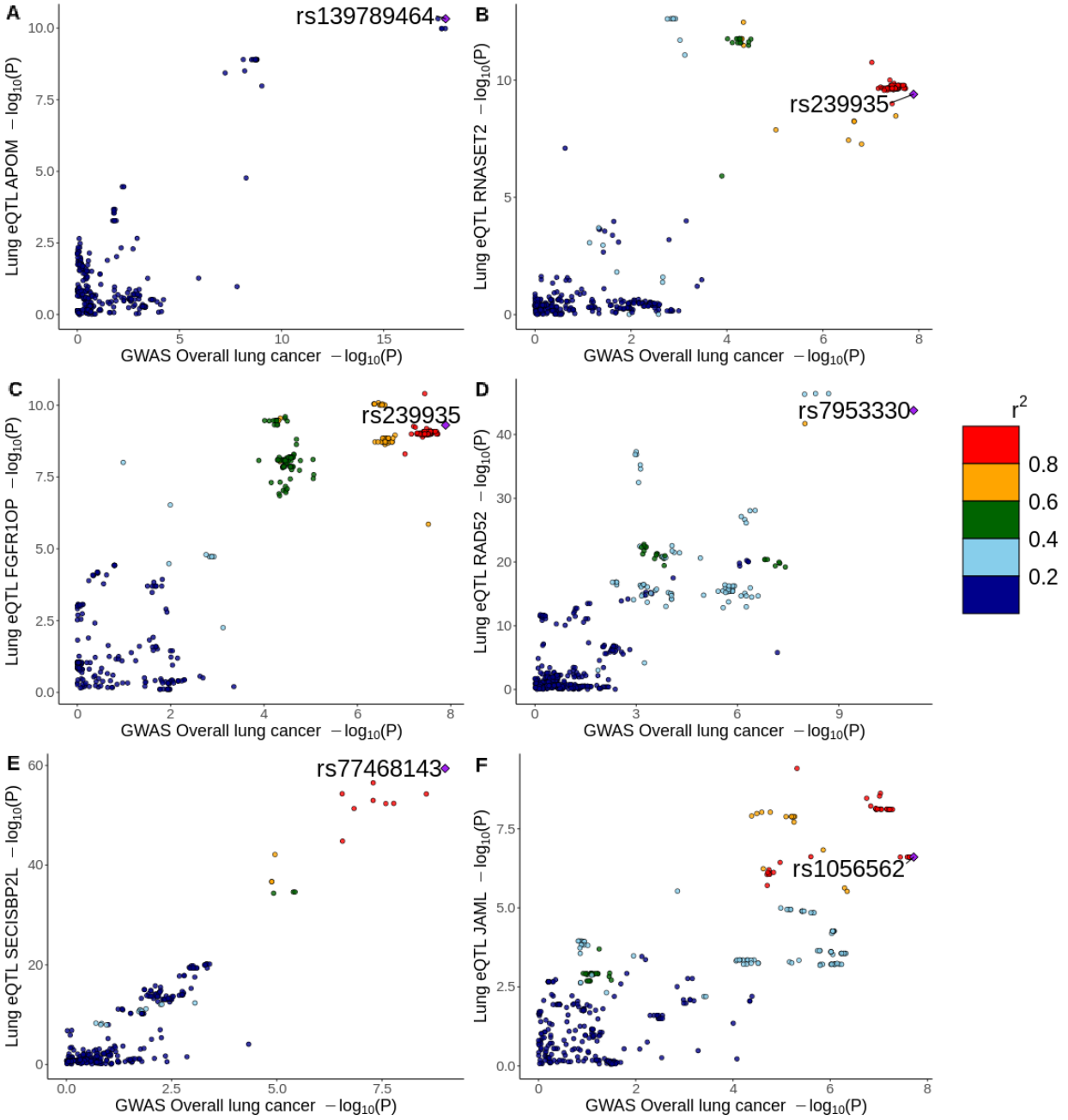
Supplementary Figure 2. LocusCompare plots for significant TWAS genes on chromosome 15q25. Association signals for SNPs within 50 Kb up and downstream of target genes are illustrated. A) *IREB2*, B) *CHRNA5*, C) *CHRNA3*, D) *HYKK*, and E) *PSMA4*.



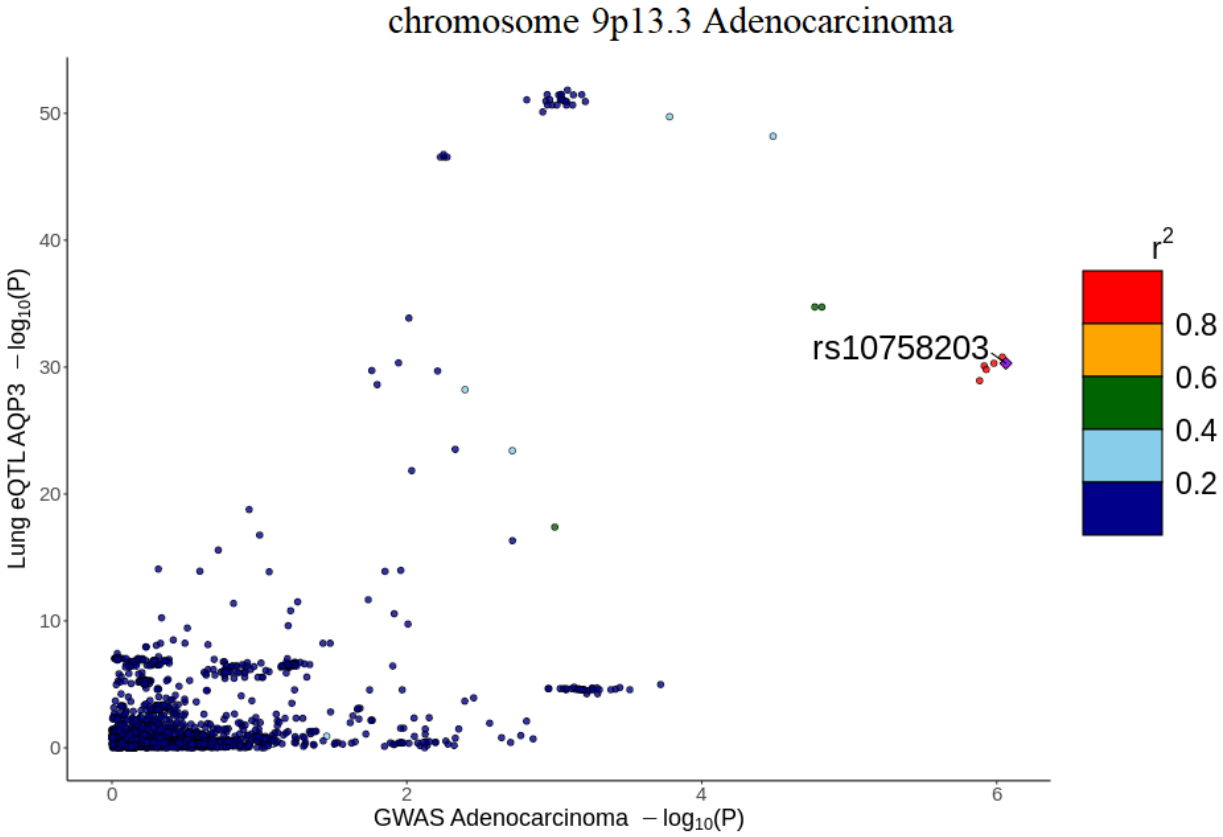
Supplementary Figure 3. TWAS results for probe sets located in the MHC locus (6p21). Results for S-PrediXcan (top) and FUSION (bottom) are illustrated in a mirror view to show similarity and differences between the two TWAS approaches. Each point represents a probe set. The x-axis shows the TWAS z-scores. The P values for gene expression-lung cancer associations are on the y-axis in -log10 scale. Annotations for the significant probe sets are indicated. The six genes in common between S-PrediXcan and FUSION with consistent direction of effect are in bold. Purple lines represent the Bonferroni significant thresholds.



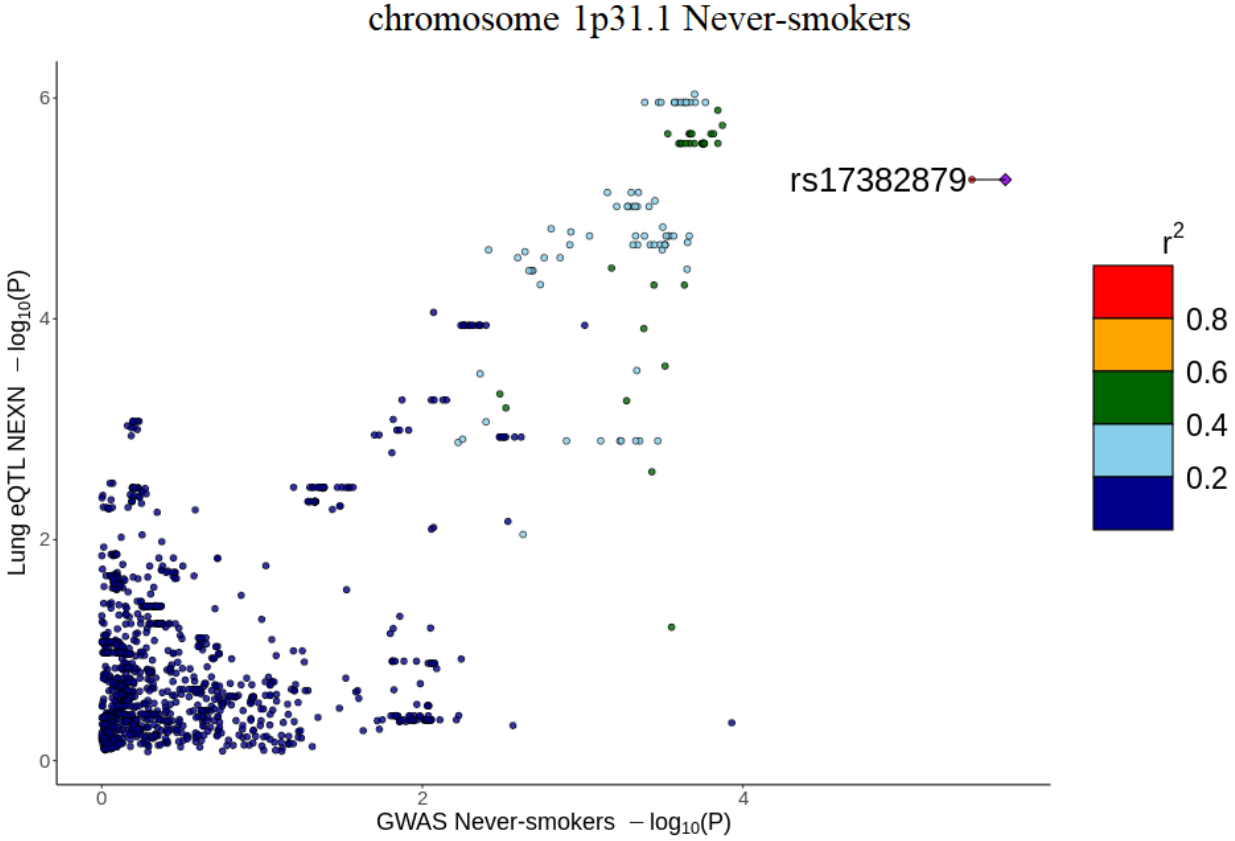
Supplementary Figure 4. LocusCompare plots for TWAS hits for overall lung cancer. Association signals for SNPs within 50 Kb up and downstream of target genes are illustrated. A) *APOM* on 6p21, B) *RNASET2* on 6q27, C) *FGFR1OP* on 6q27, D) *RAD52* on 12p13.33, E) *SECISBP2L* on 15q21.1, and F) *JAML* on 11q23.3.



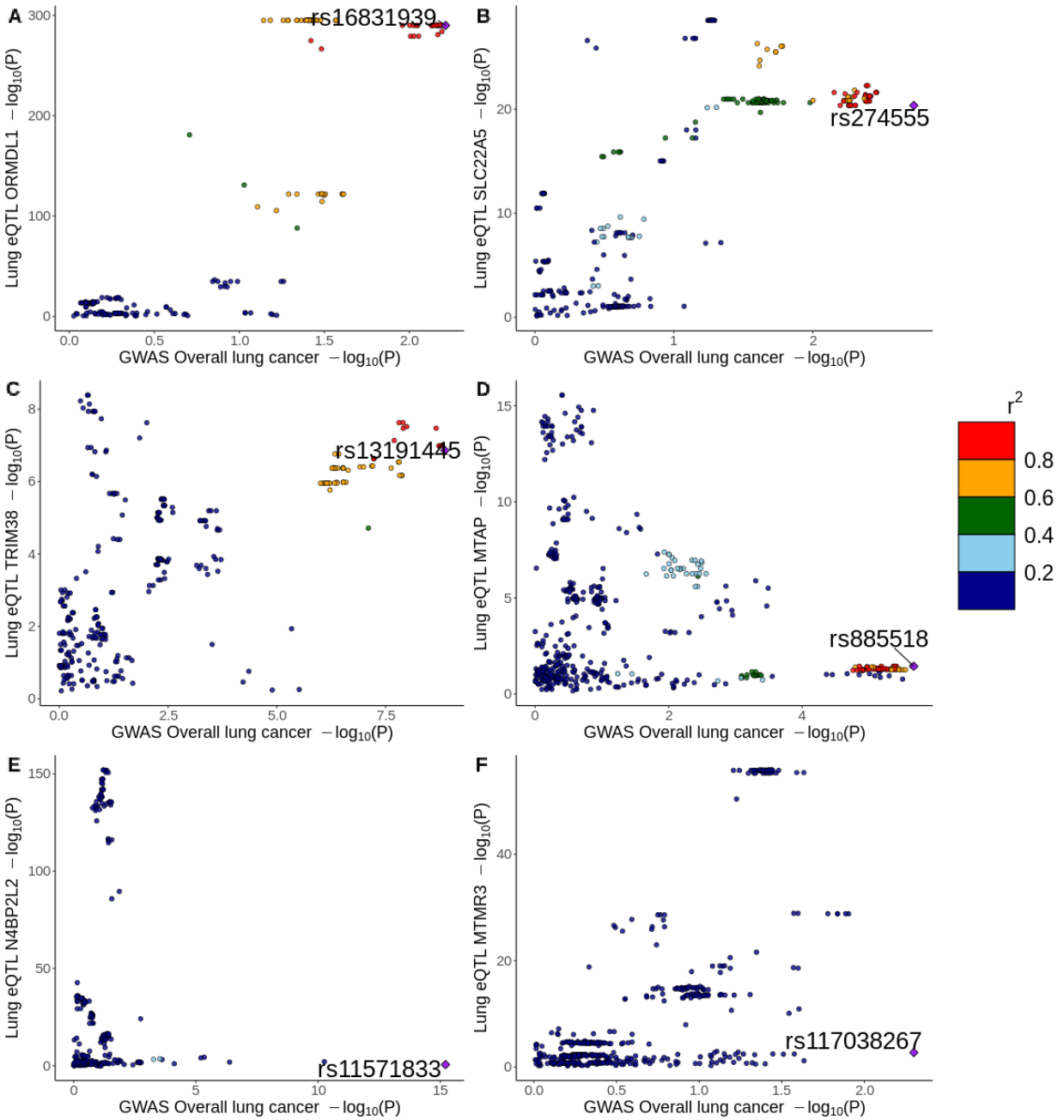
Supplementary Figure 5. LocusCompare plot for *AQP3* on chromosome 9p13.3. Association signals for SNPs within 50 Kb up and downstream of *AQP3* are illustrated. The GWAS lead variant for adenocarcinoma (rs10758203) is also strongly associated with *AQP3* expression in lung tissues.



Supplementary Figure 6. LocusCompare plot for *NEXN* on chromosome 1p31.1. Association signals for SNPs within 50 Kb up and downstream of *NEXN* are illustrated. The GWAS lead variant for lung cancer in never-smokers (rs17382879) is also associated with *NEXN* expression in lung tissues and in LD with the top eQTL-SNPs.



Supplementary Figure 7. LocusCompare plots for the top TWAS genes in known GWAS lung cancer risk loci that show $P_{TWAS}<0.05$ in both S-PrediXcan and FUSION. Association signals for SNPs within 50 Kb up and downstream of target genes are illustrated. A) *ORMDL1* on 2q32.2, B) *SLC22A5* on 5q31, C) *TRIM38* on 6p22.2, D) *MTAP* on 9p21.3, E) *N4BP2L2* on 13q13.1, and F) *MTMR3* on 22q12.2.



Supplementary Figure 8. The lung cancer gene map. The map is an ideogram of the 22 autosomal human chromosomes. The 45 lung cancer risk loci derived from published GWAS on lung cancer[1] are depicted in blue. The new adenocarcinoma susceptibility locus identified in this study (9p13.3-*AQP3*) is illustrated in red. Annotation is provided for loci for which insightful results were obtained about candidate causal genes. The colors of gene names correspond to their association with overall lung cancer (black), adenocarcinoma (blue), squamous cell carcinoma (green), and never-smokers (magenta). Bonferroni-corrected TWAS genes are in bold.